

Компьютерный практикум: стохастический анализ и моделирование

д. ф.-м. н. С.Н. Смирнов, В.А. Сливинский

20 марта 2024 г.

кафедра системного анализа, факультет ВМК, МГУ им. М.В. Ломоносова

Аннотация

Настоящее пособие ориентировано на студентов 3 и 4 курса кафедры системного анализа факультета ВМК Московского Государственного Университета им. М.В. Ломоносова. Пособие содержит типовые задания компьютерного практикума по курсу “Стохастический анализ и моделирование”, требования и общие рекомендации по выполнению практикума, полезные комментарии, образцы решений и теоретическую информацию, необходимую для успешного выполнения практикума. В приложении содержится пример отчета, выполненного студентом кафедры системного анализа.

Компьютерный практикум является неотъемлемой частью курсов “Введение в стохастический анализ” и “Стохастический анализ и моделирование”, читаемых студентам 3 и 4 курса кафедры системного анализа факультета ВМК Московского Государственного Университета им. М.В. Ломоносова. Выполнение компьютерного практикума не только позволяет повторить и закрепить полученные в рамках обучения знания из области теории вероятностей, математической статистики, стохастического анализа и математического моделирования, но и приобрести практический опыт математического моделирования, имитационного моделирования и использования метода Монте-Карло, что, безусловно, ценно для дальнейшей научной работы и профессиональной деятельности.

Авторы благодарят: А.Ю. Заночкина — за помощь в составлении практикума и разработке методических рекомендаций, Д. Сотникова — за предоставленный отчет и добросовестную работу.

Оглавление

1	Общие рекомендации по выполнению практикума	1
1.1	Сроки и правила сдачи	1
1.2	Рекомендации по оформлению программ	2
1.3	Рекомендации по оформлению отчета	4
2	Типовые задания с комментариями	7
2.1	Задание 1	7
2.1.1	Условие	7
2.1.2	Комментарии и рекомендации	7
2.2	Задание 2	9
2.2.1	Условие	9
2.2.2	Комментарии и рекомендации	9
2.3	Задание 3	12
2.3.1	Условие	12
2.3.2	Комментарии и рекомендации	12
2.4	Задание 4	14
2.4.1	Условие	14
2.4.2	Комментарии и рекомендации	14
2.5	Задание 5	15
2.5.1	Условие	15
2.5.2	Комментарии и рекомендации	15
2.6	Задание 6	16
2.6.1	Условие	16
2.6.2	Комментарии и рекомендации	16
2.7	Задание 7	17
2.7.1	Условие	17
2.7.2	Комментарии и рекомендации	17
2.8	Задание 8	19
2.8.1	Условие	19
2.8.2	Комментарии и рекомендации	19
2.9	Задание 9	21
2.9.1	Условие	21
2.9.2	Комментарии и рекомендации	21
2.10	Задание 10	22
2.10.1	Условие	22
2.10.2	Комментарии и рекомендации	22
2.11	Задание 11	24
2.11.1	Условие	24
2.11.2	Комментарии и рекомендации	24
A	Программа курса “Стохастический анализ и моделирование”	25
A.1	Программа курса	25
A.2	Рекомендуемая литература	27
A.3	Дополнительная литература	27
B	Пример выполненного отчета	29

Общие рекомендации по выполнению практикума

1.1 Сроки и правила сдачи

Выполнение практикума подразумевает реализацию студентами программ, решающих поставленные задачи, написание отчета и последующую сдачу решений и отчета проверяющему.

Задания практикума разделены на две части, каждую из которых можно сдавать независимо. Сдать нужно обе части. Каждая из частей считается выполненной, если студентом предоставлены листинги исходного кода и работающие программы, корректно решающие поставленные задачи и оформленные в соответствии с требованиями к оформлению программ (см. ниже).

Отчет является неотъемлемой частью практикума. Во время сдачи программы необходимо иметь черновую версию отчета. На момент сдачи программы, отчет должен содержать обоснования, переходы, ключевые определения и теоремы. Чистовая версия отчета сдается после сдачи решений и проверяется отдельно, результаты проверки черновой версии отчета при проверке чистовой версии не учитываются. При использовании сред Jupyter Notebook или Jupyter Lab черновая версия отчета может быть выполнена непосредственно в среде, а не в виде отдельного документа.

Сдача решений и чистовой версии отчета может производиться очно (на ПК сдающего) или дистанционно (по электронной почте). Проверяющим при выдаче практикума рекомендуется установить предельные сроки (deadline) сдачи каждой из двух частей практикума и отчета. За нарушение каждого из предельных сроков итоговая оценка сдающего понижается на 0.5 балла, при сдаче в следующем за годом выдачи календарном году — дополнительно понижается на 1 балл.

Предложенная система оценивания предполагает, что практикум должен быть выполнен в полном объеме (как неотъемлемая часть курса). С учетом интегрированности практикума в курс и его важности для формирования умений и закрепления полученных на лекциях знаний, мы настоятельно не рекомендуем отходить от настоящих рекомендаций при оценивании и выставлять удовлетворительные оценки за частично или недобросовестно выполненный практикум.

При дистанционной проверке работы, время проверки и ответа стоит вычитать из общего времени, при обнаружении ошибок, предельные сроки для данного сдающего следует продлить на время, затраченное проверяющим на проверку и ответ. Промежуточные и итоговые результаты проверки настоятельно рекомендуется публиковать в виде таблицы, доступной для просмотра как самим студентам, так и сотрудникам кафедры.

1.2 Рекомендации по оформлению программ

Мы настоятельно рекомендуем придерживаться приведенных ниже рекомендаций по оформлению программ не только для облегчения проверки работ студентов, но и для формирования навыков написания “продуктового” программного кода. Вполне вероятно, что студентам, выполняющим настоящий практикум, в обозримом будущем придется работать с программным кодом совместно, в команде. Соблюдение настоящих рекомендаций способствует выработке навыка написания читаемого, легко поддерживаемого кода.

Выполнять практикум можно на любом языке программирования из списка: Matlab, Python 3, Python 2, C, C++, Pascal. Языки Matlab и Python 3 являются предпочтительными, дальнейшие комментарии и рекомендации даются для этих языков.

Для каждого задания следует выделить отдельный скрипт, файл или функцию, внутри которого (которой) должны явно выделяться три логических блока:

1. Задание или получение входных параметров
2. Расчет. Вызов соответствующих функций, элементарные операции над результатами этих функций.
3. Вывод результата и/или графика (графиков).

Если целью задания является реализация генератора некоего распределения, то по умолчанию из библиотечных генераторов разрешено пользоваться только генератором равномерно распределенных случайных величин на интервале $[0, 1]$, а также самостоятельно реализованными ранее генераторами. При наличии в языке сторонних библиотек, которые предоставляют реализации критериев для проверки статистических гипотез (например, `scipy` для языка Python), не допускается использование библиотечных критериев во всех заданиях.

Код должен быть оригинальным и уникальным, с соблюдением единого авторского стиля. Использование чужого кода и/или кода из общедоступных источников не допускается. Переменные, методы (функции) и иные части программы должны иметь понятные и интерпретируемые (но, по возможности, не слишком длинные) названия на английском языке. Недопустимо использование транслита: `setka` вместо `grid`, `zadanie` вместо `task`, `vyborka` вместо `sample` и тому подобное.

В качестве примера: плохим названием для метода, генерирующего сетку для численного эксперимента, будет `generate_grid_for_numerical_experiment` (поскольку оно слишком длинное), так и `ggfnn` (поскольку оно абсолютно неинтерпретируемое), равно как и `poluchi_setku` (поскольку оно использует транслит). Подобный метод коротко и понятно можно назвать `get_grid`.

В целом, рекомендуется сохранять в названиях лаконичность и однородность: использовать префиксы `get` и `set` для возвращающих и устанавливающих значения методов соответственно, общепринятые сокращения `rand` (`random`), `dist` (`distribution`),

gen (generate).

В именах индексов и счётчиков цикла допустимо использование последовательных букв i , j , k . Стоит обратить внимание, что при наличии вложенных циклов, следует сохранять последовательность индексов (для трижды вложенного цикла разумно ожидать, что i отвечает внешнему циклу, j — среднему, а k — внутреннему).

Особое внимание стоит уделить правописанию иностранных фамилий и имен, например:

- фон Нейман — von Neumann
- Коши — Cauchy
- Орнштейн — Ornstein
- Уленбек — Uhlenbeck
- Декартова — Cartesian

Крайне вредной привычкой является использование в программе т.н. «магических чисел»: по возможности, следует избегать использования любых цифр, кроме 0 и 1 везде, кроме присвоения значения переменной.

Скажем, для цикла на 1234 шага, стоит завести отдельную переменную (например, $n = 1234$), после чего проводить цикл по i от 1 до n .

Компактность и читаемость. Код функции или скрипта должен уместиться на одну страницу. Регулярно повторяющиеся блоки должны быть вынесены в отдельные функции. Расчеты и визуализация внутри исходного кода должны быть разделены, каждый блок расчетов должен быть выделен в отдельную функцию с соответствующими входными параметрами и выходными значениями.

Обязательным является использование векторных возможностей языка. Число циклов должно быть минимальным, алгоритм решения должен быть оптимальным или близок к оптимальному.

Следует минимизировать используемый объем памяти. При генерации больших выборок стоит разделить задачу на менее объемные и выполнить расчет в цикле (по «эпохам»). Например, расчет среднего для n независимых одинаково распределенных случайных величин с генератором $\text{genIID}(n)$ можно реализовать как $\text{sum}(\text{genIID}(n))$. При значениях $n > 10^8$ памяти может не хватить, в этом случае лучше использовать циклический вызов функции на меньших объемах выборки, а результат представлять в виде итоговой суммы отдельных блоков.

При визуализации работы датчиков дискретных распределений не рекомендуется использование функций, которые строят непрерывные линии. Рекомендуется пользоваться функциями гистограмм и столбчатых диаграмм (bar'ов).

Продемонстрировать свойства отсутствия памяти у распределений необходимо любым из двух способов:

1. При помощи наложения соответствующих гистограмм двух выборок, без смещения одной относительно другой. Столбцы на графиках в пределе должны накладываться и совпасть. При этом столбцы, находящиеся на переднем плане, должны иметь меньшую толщину (либо быть прозрачными), для того, чтобы не перекрывать полностью задний план.
2. При помощи наложения соответствующих эмпирических функций распределения. Аналогично, в пределе графики должны совпадать, а визуализация не должна содержать перекрытия графиками друг друга.

При демонстрации гистограмм различных распределений рекомендуется аналитически определить диапазон распределения, где сосредоточено более $1 - \alpha \approx 99.9\%$ выборки, и соответствующим образом ограничить оси (функцией `xlim` или аналогичными).

1. Так, к примеру, для нормального распределения этот диапазон часто определяют из “правила трех сигм”, то есть при известных параметрах μ, σ наблюдаемого распределения разумно ограничить график по оси x значениями $\mu \pm 3 \cdot \sigma$. В действительности, это правило применимо и для многих других распределений, у которых значения среднего и дисперсии (стандартного отклонения) могут быть легко найдены аналитически.
2. Для таких распределений, как биномиальное, пуассоновское и других, которые можно приблизительно считать унимодальными и симметрическими, также применимо “правило трех сигм”.
3. При демонстрации гистограммы геометрического или экспоненциального распределения разумно воспользоваться знанием соответствующей функции распределения $F(\cdot)$ и ограничить ось абсцисс снизу нулем, а сверху — значением $F^{-1}(1 - \alpha)$.

1.3 Рекомендации по оформлению отчета

Оформление отчета должно быть выполнено в системе *L^AT_EX* с соблюдением общеуниверситетских, общефакультетских и кафедральных требований к оформлению отчетов.

Отчеты должны быть уникальными: формулировки теорем и определения в разных отчетах могут совпадать, но пояснения к формулам и их описание должны быть оригинальными. Тем более, оригинальными должны быть графики и значения измерений.

Каждое задание следует иллюстрировать графически с объяснением полученных результатов. Графики должны быть оформлены в одном стиле, все обозначения должны быть видны. Необходимо подобрать универсальное разрешение (или несколько разрешений) на графиках. Для этого можно сохранять графики при помощи скрипта или в командной строке с фиксированными настройками. Все приведенные графики должны иметь подписанные оси, все изображенные на графиках объекты должны быть обозначены в легенде или в подписи. Для языка Matlab, команду `title` рекомендуется не использовать, `grid on` или `grid minor` — желательно включать.

При оформлении заданий, в которых используются статистические критерии, рекомендуется произвести достаточно много независимых испытаний и указать:

1. Количество испытаний
2. Уровень значимости
3. Частоту принятия соответствующей гипотезы

Отчет должен быть написан на грамотном русском языке и не содержать грамматических, орфографических, стилистических и синтаксических ошибок. При большом числе ошибок, отчет может быть возвращен на доработку без дополнительного указания ошибок по существу.

Требуется обосновывать все используемые построения, в том числе, ссылками на соответствующие теоремы и факты. При использовании нетривиальных теорем необходимо привести их полную формулировку с доказательством или ссылкой на литературу.

Листинг исходного кода (исходный код программ) в отчет включать не требуется.

Типовые задания с комментариями

2.1 Задание 1

2.1.1 Условие

1. Реализовать генератор схемы Бернулли с заданной вероятностью успеха p . На основе генератора схемы Бернулли построить датчик биномиального распределения.
2. Реализовать генератор геометрического распределения; проверить для данного распределения свойство отсутствия памяти.
3. Промоделировать игру в орлянку: бесконечную последовательность независимых испытаний Бернулли с бросанием “правильной” (честной, $p = 0.5$) монеты. Величина “выигрыша” S_n определяется как сумма по n испытаниям значений 1 и -1 в зависимости от выпавшей стороны монеты. Проиллюстрировать в виде ломаной поведение нормированной суммы $Y(i) = \frac{S_i}{\sqrt{n}}$ как функцию от номера испытания i для отдельно взятой траектории. Дать теоретическую оценку для значения $Y(n)$ при $n \rightarrow \infty$

2.1.2 Комментарии и рекомендации

Напомним, что схемой Бернулли с заданной вероятностью p называется эксперимент, состоящий из серии испытаний, удовлетворяющих следующим условиям:

- Отсутствие взаимного влияния в испытаниях;
- Воспроизводимость. Однородные испытания проводятся в сходных (не обязательно идентичных) условиях;
- Существует признак, который реализуется (“успех” с вероятностью p) или не реализуется (“неуспех” с вероятностью $q = (1 - p)$) в испытании. Признак может быть отнесен к любому из испытаний (в силу их однородности).

Здесь и далее, под генератором и датчиком подразумевается функция, возвращающая выборку заданного размера из соответствующего распределения. Таким образом, датчик из первого задания должен принимать в качестве входных аргументов не только вероятность успеха p , но и параметры размера возвращаемого вектора/матрицы (как минимум, два числа m, n для случая вектора и матрицы). В последующих заданиях следует пользоваться как самим генератором из данного задания, так и возможностью генерации вектора / матрицы (для того, чтобы исключить лишние циклы и вызовы генератора).

Выполняя задание, следует помнить, что входные параметры нужно проверять на “адекватность” и соответствие логике. К примеру, параметр p (вероятность успеха) вполне может быть задан большим единицы или меньшим нуля. В таком случае, возможно несколько равнозначных вариантов поведения:

1. Программа может выдавать ошибку и явно требовать нахождения параметра p в интервале $[0, 1]$.
2. Программа может зажимать значение p в интервал $[0, 1]$ (методы `clamp` и аналогичные).

При реализации генератора геометрического распределения, следует внимательно отнестись к тому, какое именно определение геометрического распределения используется. В генераторе геометрического распределения не допускается использование циклов со счетчиком (циклов `for`), допускается использование не более одного цикла с предусловием (цикла `while`).

Пример реализации генератора схемы Бернулли на языке Matlab:

```
function x = bernrnd(p,varargin)
if (p>1) || (p<0)
    error('p = %4.2f is out of bounds [0, 1]',p)
end
x = rand(varargin{:}) < p;
end
```

При исследовании нормированных сумм в третьей части задания следует воспользоваться центральной предельной теоремой.

2.2 Задание 2

2.2.1 Условие

1. Построить датчик сингулярного распределения, имеющий в качестве функции распределения канторову лестницу. С помощью критерия Колмогорова убедиться в корректности работы датчика.
2. Для канторовых случайных величин с помощью критерия Смирнова проверить свойство симметричности относительно $\frac{1}{2}$ (X и $1 - X$ распределены одинаково) и свойство самоподобия относительно деления на 3 (условное распределение Y при условии $Y \in [0, 1/3]$ совпадает с распределением $\frac{Y}{3}$).
3. Рассчитать значения математического ожидания и дисперсии для данного распределения. Сравнить теоретические значения с эмпирическими (для различных объемов выборок), проиллюстрировать сходимость эмпирических значений к теоретическим.

2.2.2 Комментарии и рекомендации

Под канторовой лестницей понимается непрерывная монотонная функция $F : [0, 1] \rightarrow [0, 1]$, построенная следующим образом:

- $F(0) = 0, F(1) = 1$
- Разбиваем отрезок $[0, 1]$ на три равные части. Внутри центральной части полагаем $F(x) = \frac{1}{2}, x \in [\frac{1}{3}, \frac{2}{3}]$.
- Два оставшихся сегмента снова разбиваем на три равные части, полагая в центральных сегментах $F(x) = \frac{1}{4}$ и $\frac{3}{4}$.
- Продолжаем данный процесс до бесконечности, определяя на внутренних сегментах $F(x)$ равной среднему арифметическому из соседних определенных значений. На остальных точках отрезка $[0, 1]$ определяем $F(x)$ по непрерывности.

Полученная функция может быть также определена через двоичную и троичную запись числа $x \in [0, 1]$: для этого, в троичной записи числа x отбросим все цифры после первой (самой левой) единицы, а в оставшейся записи заменим все двойки на 1. Полученная последовательность даст запись значения канторовой лестницы в точке x в двоичной системе счисления.

Пример реализации генератора случайной величины с неким сингулярным распределением, который помимо самих значений случайной величины рассчитывает значение функции распределения (на языке Matlab):

```
function [x,F] = cantrnd(n,eps)
```

```

if nargin < 1
    n = 1; % объем выборки
end
if nargin < 2
    eps = 1e-10; % точность расчета
end
m = round(-log(eps)/log(9));
bern = bernrnd(0.5,n,m);
deg = -(1:m)';
x = 6 * bern * 9.^deg;
F = bern * 2.^deg;
end

```

Для расчета статистик Колмогорова и Смирнова для заданных выборок известного размера достаточно использовать следующие функции (или их аналоги):

- `sort()` — сортировка вектора/массива
- `abs()` — вычисление абсолютного значения (модуля)
- `max()` — вычисление максимального значения
- `sqrt()` — извлечение квадратного корня
- `cumsum()` — кумулятивная сумма вектора/массива (результатом которой для массива $x = \{x_1, \dots, x_n\}$ является массив значений $s = \left\{x_1, x_1 + x_2, \dots, \sum_{i=1}^n x_i\right\}$)

Для проверки простой гипотезы H_0 о соответствии выборки известному закону распределения с заданным уровнем значимости α и правосторонней критической областью можно воспользоваться одним из двух подходов:

- Подход на основе p-значения (p-value) для статистики $\sqrt{n}D_n$, где D_n — статистика Колмогорова. Нулевая гипотеза принимается, если p-value $p = 1 - F_K(\sqrt{n}D_n)$ превышает заданный уровень значимости α . Для расчёта p-value требуется реализовать собственную функцию, возвращающую значение функции распределения Колмогорова F_K в заданной точке, для чего можно воспользоваться следующим представлением функции распределения Колмогорова:

$$F_K(x) = 1 + 2 \sum_{k=1}^{+\infty} (-1)^k e^{-2k^2 x^2}$$

- Подход на основе критического значения (верхнего квантиля распределения). Гипотеза о соответствии выборки принимается с заданным уровнем значимости α , если $\sqrt{n}D_n < K_\alpha$, где $K_\alpha = F_K^{-1}(1 - \alpha)$. Этот подход требует обращения функции распределения, что аналитически невозможно. Для $\alpha < 50\%$ можно воспользоваться легко обратимым приближением функции распределения Колмогорова $F_K(x) \approx 1 - 2e^{-2x^2}$. Для $\alpha > 50\%$ можно воспользоваться известными табличными значениями квантилей распределения Колмогорова.

Для практического подтверждения найденных значений среднего и дисперсии необходимо построить графики сходимости выборочного среднего и дисперсии к их теоретическим значениям без использования циклов.

По критерию Колмогорова решение принимается на основе двух векторов:

1. Вектора значений выборки;
2. Вектора значений функции распределения в этих точках выборки.

Следовательно, функция, осуществляющая реализацию критерия Колмогорова, должна принимать на вход два вектора, а выдавать либо p -value, либо значение статистики, либо (в случае, если уровень значимости α является параметром функции) логическое значение, отвечающее принятию или отвержению гипотезы.

Критерий Смирнова принимает решение только на основе двух векторов (возможно разной длины):

1. Вектора значений первой выборки;
2. Вектора значений второй выборки.

Для принятия решения по этому критерию совсем не обязательно знать закон распределения, к которому (предположительно) относятся выборки.

Следует также помнить, что критерий Колмогорова и критерий Смирнова имеют правостороннюю критическую область.

2.3 Задание 3

2.3.1 Условие

1. Построить датчик экспоненциального распределения. Проверить для данного распределения свойство отсутствия памяти.
2. Пусть X_1, \dots, X_n — независимые экспоненциально распределенные случайные величины с параметрами $\lambda_1, \dots, \lambda_n$. Найти распределение случайной величины $Y = \min(X_1, \dots, X_n)$.
3. На основе датчика экспоненциального распределения построить датчик пуассоновского распределения.
4. Построить датчик пуассоновского распределения как предел биномиального распределения. Убедиться в корректности построенного датчика при помощи критерия χ^2 Пирсона.
5. Построить датчик стандартного нормального распределения методом моделирования случайных величин парами с переходом в полярные координаты (преобразование Бокса-Мюллера). Проверить при помощи t-критерия Стьюдента равенство математических ожиданий, а при помощи критерия Фишера — равенство дисперсий.

2.3.2 Комментарии и рекомендации

В первом пункте задания, следует не только найти теоретический закон распределения случайной величины Y , но и подтвердить его эмпирически.

При моделировании экспоненциального распределения разумно воспользоваться методом обращения функции распределения и теоретически обосновать корректность построенного датчика.

Свойство отсутствия памяти для непрерывного распределения может быть сформулировано аналогично дискретному случаю:

$$\mathbb{P}(X > m + n | X > m) = \mathbb{P}(X > n), \forall m, n > 0$$

При визуализации свойства отсутствия памяти следует помнить, что построенные графики/гистограммы не должны перекрывать друг друга.

При использовании различных критериев, стоит обратить свое внимание на то, что они имеют разные критические области:

- критерий согласия Пирсона имеет правостороннюю критическую область;

- критерий Фишера имеет двустороннюю критическую область;
- t-критерий Стьюдента имеет двустороннюю критическую область.

В работе и отчете следует обязательно разобраться с тем, как тип критической области влияет на способ принятия гипотезы.

При моделировании стандартных нормальных величин методом Бокса-Мюллера обратите внимание на то, какую версию вы используете: одна из версий метода является более быстрой, но требует использования цикла с предусловием или ветвления.

2.4 Задание 4

2.4.1 Условие

1. Построить датчик распределения Коши.
2. На основе датчика распределения Коши с помощью метода фон Неймана построить датчик стандартного нормального распределения. При помощи графика `normal probability plot` убедиться в корректности построенного датчика и обосновать наблюдаемую линейную зависимость.
3. Сравнить скорость моделирования стандартного нормального распределения в задании 3 и в задании 4.

2.4.2 Комментарии и рекомендации

Метод фон Неймана также известен как метод элиминации и как `rejection sampling`. В отчете стоит не только дать описание метода, но и объяснить, почему он действительно работает и обосновать корректность.

С учетом “случайности” метода (и возможности отклонения значений), программа не может обойтись без цикла с предусловием. Цикл со счетчиком использовать крайне нежелательно.

Рекомендуется самостоятельно разобраться и ответить в отчете на следующие вопросы, касающиеся графика `normal probability plot`:

1. Каким образом он задаётся и определяется?
2. Почему он будет линейным не только для выборки из стандартного нормального распределения, но и для выборки из $\mathcal{N}(\mu, \sigma^2)$? Как связаны наклон и сдвиг наблюдаемой кривой с параметрами μ, σ ?
3. Почему при построении по выборке из любого другого распределения кроме нормального график не будет линейным?
4. Для каких еще классов распределений графики `normal probability plot` будут инварианты относительно параметров масштаба и сдвига?

2.5 Задание 5

2.5.1 Условие

1. Пусть $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Убедиться эмпирически в справедливости теоремы о законе больших чисел (ЗБЧ) и центральной предельной теоремы (ЦПТ): исследовать поведение суммы $S_n = \sum_{i=1}^n X_i$ и эмпирического распределения величины

$$\sqrt{n} \frac{S_n}{n} - a$$

2. Считая μ, σ неизвестными, построить доверительные интервалы для среднего и дисперсии по имеющейся выборке.
3. Пусть $X_i \sim K(a, b)$ — имеет распределение Коши с параметром сдвига a и масштаба b . Изучить эмпирически как ведут себя суммы $\frac{S_n}{n}$, объяснить результат и найти закон распределения данных сумм.

2.5.2 Комментарии и рекомендации

В данном задании следует пользоваться любым из ранее построенных генераторов нормальных случайных величин. Для пунктов 1 и 2 стоит сгенерировать разные выборки. Для демонстрации работы ЗБЧ достаточно одной выборки достаточно большой длины, после чего посчитать накопленное выборочное среднее и проиллюстрировать его эволюцию. При демонстрации работы ЦПТ необходимо построить несколько выборок фиксированной одинаковой длины и продемонстрировать сходимость соответствующих статистик от этих выборок.

При построении доверительных интервалов настоятельно рекомендуется изобразить их эволюцию и отобразить теоретические значения на том же графике. При этом, при использовании большой выборки, допускается некоторое прореживание.

2.6 Задание 6

2.6.1 Условие

1. Вычислить следующий интеграл:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \frac{e^{-(x_1^2 + \cdots + x_{10}^2 + \frac{1}{2^7 \cdot x_1^2 \cdots x_{10}^2})}}{x_1^2 \cdots x_{10}^2} dx_1 dx_2 \cdots dx_{10}$$

2. Методом Монте-Карло

3. Методом квадратур, сводя задачу к вычислению собственного интеграла Римана.

2. Оценить точность вычислений для каждого из двух случаев.

2.6.2 Комментарии и рекомендации

В данном задании допускается использование библиотечного генератора нормальных случайных величин.

Для демонстрации работы метода Монте-Карло необходимо включить в отчет таблицу с результатами независимых испытаний, а также оцененной точностью результатов. При этом, настоятельно рекомендуется использовать и другие визуальные методы, иллюстрирующие скорость расчетов и точность результатов.

Для вычисления методом квадратур, очевидно, потребуется произвести замену переменной. При этом, как часто бывает при решении сложных численных задач, некоторые свойства рассматриваемого интеграла и подынтегральной функции могут существенно упростить расчеты без потери точности вычислений.

При вычислении методом квадратур может возникнуть соблазн использовать несколько вложенных циклов. Ни в коем случае не стоит поддаваться ему и писать десять вложенных циклов: подобный программный код эквивалентен грубым грамматическим и лексическим ошибкам в докторской диссертации. Как минимум, стоит сгруппировать индексы (координаты узлов сетки, на которой выполняется расчет) в массив и производить итерации по элементам этого массива. Для формирования сетки полезной может оказаться функция, реализующая декартово произведение массивов (Cartesian product): можно воспользоваться встроенными функциями (такие функции присутствуют, например, в языках Matlab и Python), или реализовать свою собственную.

В отчёте должны присутствовать теоретические оценки точности для двух методов и их обоснования.

2.7 Задание 7

2.7.1 Условие

1. Методом случайного поиска найти минимальное значение функции f на множестве $A = \{x_1, x_2 : x_1^2 + x_2^2 \leq 1\}$, где

$$f(x_1, x_2) = x_1^3 \sin \frac{1}{x_1} + 10x_1x_2^4 \cos \frac{1}{x_2}, \quad x_1, x_2 \neq 0$$

При $x_1 = 0$ или $x_2 = 0$ функция доопределяется по непрерывности.

2. Методом имитации отжига найти минимальное значение функции Розенброка g в пространстве \mathbb{R}^2 , где

$$g(x) = (x_1 - 1)^2 + 100(x_2 - x_1^2)^2$$

3. Оценить точность и сравнить результаты со стандартными методами оптимизации.

2.7.2 Комментарии и рекомендации

Как и во многих других заданиях и численных проблемах, в первом пункте разумно провести предварительный анализ функции: аналитические выводы могут существенно упростить задачу минимизации.

Метод имитации отжига также известен как *simulated annealing*. Проводя аналогию между значением функции и общей энергией системы, метод может быть описан на метаязыке следующим образом:

1. Выберем произвольное начальное состояние системы и достаточно большую “температуру” T .
2. На каждом шаге выберем “соседа” для текущего состояния системы; с некоторой функцией вероятности, зависящей от разницы значений функции в текущем положении и соседнем (разности “энергий”) и текущей температуры, осуществим переход в соседнее состояние.
3. Понизим “температуру” по некоторому закону “охлаждения”.

Таким образом, для реализации метода имитации отжига необходимо задать следующие элементы:

- Метод генерации соседа как функцию от текущего положения системы $s_i = (x_i, y_i)$ и температуры T_i .

- Функцию вероятности перехода, зависящую от текущей температуры T_i и разности значений минимизируемой функции $\Delta F = F(s_{i+1}) - F(s_i)$
- Функцию понижения температуры системы $T_{i+1} = \mathcal{T}(T_i)$

При этом, все эти функции являются параметрическими и вся сложность подбора сводится к отысканию оптимальных значений параметров для этих функций.

В качестве примера можно взять следующие семейства функций:

- Функция выбора соседа — нормальная случайная величина со средним (x_i, y_i) и дисперсией $\sigma^2 * T_i$
- Функция вероятности перехода $P = e^{-\Delta F_i/T_i}$
- Функция понижения температуры $T_{i+1} = kT_i$

При различных значениях k, σ, T_0 скорость сходимости, очевидно, будет различаться. Кроме того, при подборе параметров следует обратить внимание на стабильность алгоритма (как часто происходят “фатальные промахи”), среднее число итераций для достаточно хорошей сходимости к минимуму и число переходов в другие точки (при P близком к 0 или 1 переходы будут не происходить вообще или происходить всегда, что плохо для работы метода).

Стоит также попробовать поэкспериментировать с другими видами функций и подобрать для них оптимальные параметры. При оценке точности стоит привести эмпирические оценки и сослаться на соответствующую литературу.

2.8 Задание 8

2.8.1 Условие

1. Применить метод Монте-Карло к решению первой краевой задачи для двумерного уравнения Лапласа в единичном круге

$$\begin{cases} \Delta u = 0, (x, y) \in D \\ u|_{\delta D} = f(x, y) \\ u \in C^2(D), f \in C(\delta D) \\ D = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 \leq 1\} \end{cases}$$

2. Для функции $f(x, y) = x^2 - y^2$ найти аналитическое решение и сравнить с полученным по методу Монте-Карло.

2.8.2 Комментарии и рекомендации

Напомним, что метод Монте-Карло для данной задачи сводится к моделированию случайного блуждания частицы по некоторой сетке с переносом температуры частицы с границы в узлы.

Настоятельно рекомендуется реализовывать код с единственным циклом.

В прямоугольной системе координат алгоритм может иметь следующую структуру

Инициализация pos на границе.

for

 Случайно выбрать направление движения direc.

 if текущая точка находится за или на границе круга

 Пересмотреть текущую температуру temp на границе.

 Пересмотреть направление движения direc, если это необходимо.

 end

 Запомнить, что процесс побывал в текущей точке pos с текущей температурой temp.

 Сдвинуть текущее положение pos в выбранном случайном направлении direc.

end

Рекомендуется использовать двумерный целочисленный вектор для хранения текущей позиции процесса.

В полярных координатах алгоритм немного усложняется тем, что переходные вероятности из соседних точек не равны, то есть $P_{r,\phi}(r + \Delta r, \phi) \neq P_{r+\Delta r,\phi}(r, \phi)$. Поэтому при моделировании случайного блуждания по полярной сетке по достижению границы необходимо запомнить температуру на границе и пройти в обратном времени по траектории до предыдущего столкновения с границей, распространяя сохраненную температуру. То есть движение частицы описывается вероятностями перехода

“вперед”, а температура распространяется “назад”. Пример псевдокода для полярной сетки:

Инициализация pos на границе.

for

 Случайно выбрать направление движения direc.

 if текущая точка находится на границе круга, либо в центре круга

 Пересмотреть текущую температуру temp на границе или в центре.

 Двигаться по траектории назад, распространяя текущую температуру temp.

 Пересмотреть направление движения direc.

 end

 Запомнить текущее положение pos в траектории.

 Сдвинуть текущее положение pos в выбранном случайном направлении direc.

end

2.9 Задание 9

2.9.1 Условие

Рассмотреть два вида гауссовских процессов:

- Винеровский процесс $W(t), t \in [0; 1], W(0) = 0$
- Процесс Орнштейна–Уленбека $X(t), t \in [0; 1], X(0) = X_0$, т.е. стационарный марковский гауссовский процесс. Начальные значения X_0 следует выбирать случайным образом так, чтобы полученный процесс был стационарным.

Для данных процессов:

1. Найти ковариационную функцию и переходные вероятности.
2. Про моделировать независимые траектории процесса с данными переходными вероятностями методом добавления разбиения отрезка.
3. Построить график траектории, не соединяя точки ломаной, с целью получения визуально непрерывной линии.

2.9.2 Комментарии и рекомендации

Стационарный процесс Орнштейна–Уленбека вводился как единственный нетривиальный гауссовский процесс, являющийся стационарным и обладающий марковским свойством. Именно из этих двух свойств нужно вывести вид ковариационной функции данного процесса

$$R(t_1, t_2) = \text{Cov}(X(t_1), X(t_2)) = \sigma^2 e^{-\theta|t_2-t_1|}.$$

2.10 Задание 10

2.10.1 Условие

Произвести фильтрацию одномерного процесса Орнштейна-Уленбека:

1. Используя генератор белого шума, добавить к реализации процесса Орнштейна-Уленбека случайную ошибку с заранее известной дисперсией.
2. При помощи одномерного фильтра Калмана оценить траекторию процесса по зашумленному сигналу, считая известными параметры шума и процесса.
3. Рассмотреть следующие виды шума:
 1. Гауссов
 2. Коши (шум имеет распределение Коши)

2.10.2 Комментарии и рекомендации

В задании следует рассмотреть дискретный одномерный фильтр Калмана для динамической системы вида:

$$\begin{aligned} x_{n+1} &= a \cdot x_n + \nu_n, \quad i.i.d. \nu_n \sim \mathcal{N}(0, q), \quad x_1 \sim \mathcal{N}(0, \sigma^2) \\ y_n &= x_n + \varepsilon_n, \quad i.i.d. \varepsilon_n \sim \mathcal{N}(0, r) \end{aligned}$$

Значение дисперсии шума r считается известным еще на этапе генерации этого шума.

Для определения коэффициентов уравнения динамики необходимо приравнять теоретические значения ковариационной матрицы процесса Орнштейна-Уленбека в точках t_n и t_{n+1} с соответствующими значениями динамической системы, то есть решить систему уравнений относительно неизвестных коэффициентов a и q через известные σ и θ :

$$\begin{cases} \sigma^2 = R(t_n, t_n) = \text{Var}(x_n) \\ \sigma^2 e^{-\theta(t_{n+1}-t_n)} = R(t_n, t_{n+1}) = \text{Cov}(x_n, x_{n+1}) = a \cdot \text{Var}(x_n) \\ \sigma^2 = R(t_{n+1}, t_{n+1}) = \text{Var}(x_{n+1}) = a^2 \cdot \text{Var}(x_n) + q \end{cases}$$

Напоминаем, что при фильтрации значение оригинального процесса x_n неизвестно. Вместо него доступен наблюдаемый зашумленный сигнал y_n .

Обратный проход фильтра проводить не нужно.

При визуализации работы фильтра Калмана рекомендуется изобразить:

1. Траекторию x_n процесса Орнштейна-Уленбека без шума в виде непрерывной линии.
2. Траекторию y_n зашумленного процесса в виде отдельных точек.
3. Доверительный интервал $[\hat{x}_n - k_\alpha R_n; \hat{x}_n + k_\alpha R_n]$ с уровнем значимости α для процесса Орнштейна-Уленбека в виде монотонно закрашенной области, как результат работы фильтра.

2.11 Задание 11

2.11.1 Условие

Построить двумерное пуассоновское поле, отвечающее сложному пуассоновскому процессу:

1. Система массового обслуживания. Первая координата поля — время поступления заявки в СМО (распределенное равномерно), а вторая — время обслуживания заявки (распределение χ^2 с десятью степенями свободы).
2. Система массового обслуживания с циклической интенсивностью $\lambda(1 + \cos(t))$ и единичными скачками. При помощи метода Льюиса и Шедлеара, свести задачу моделирования неоднородного пуассоновского процесса к моделированию двумерного пуассоновского поля, где первая координата распределена равномерно, а вторая имеет распределение Бернулли.
3. Работа страховой компании: первая координата — момент наступления страхового случая (равномерное распределение), вторая — величина ущерба (распределение Парето). Поступление капитала считать линейным по времени со скоростью $c > 0$, начальный капитал $W > 0$.

2.11.2 Комментарии и рекомендации

Стоит помнить, что в третьем пункте капитал не может быть отрицательным (т.е. имеет место ограничение типа неравенства на капитал).

Для каждой из систем настоятельно рекомендуется изучить и рассмотреть все возможные принципиально разные случаи поведения системы (для разных значений параметров), аналитически найти критические значения параметров, при переходе через которые поведение системы меняется существенно, обосновать эти изменения.

Приложение А

Программа курса “Стохастический анализ и моделирование”

Предполагается, что слушатель знаком с основами теории вероятностей. Одной из задач курса является систематизация знаний на основе теоретического фундамента – аксиоматики Колмогорова (на базе теории меры), с особым фокусом на интерпретацию основных понятий, необходимую для приложений. Другая цель – анализ часто используемых на практике вероятностных моделей, а также изложение основ метода статистического моделирования (метода Монте-Карло). Составной частью курса является компьютерный практикум, иллюстрирующий курс и развивающий у студентов технику статистического моделирования, в том числе для стохастических процессов.

А.1 Программа курса

1. Математическая модель явлений, в которых случайность выступает существенным фактором. Аксиоматика Колмогорова и непротиворечивость модели; терминология, содержание и интерпретация объектов. Объективно-частотная и субъективная интерпретации вероятности. Природа случайности. Тезис Пуанкаре о неустойчивости.
2. Интерпретация условной вероятности, аксиома связь априорной и апостериорной модели, объективно-частотное обоснование. Парадокс фон Мизеса. Независимость событий. Схема Бернулли как простейшая содержательная модель. Парадокс Д’Аламбера.
3. Алгоритмический подход Колмогорова. Универсальный фильтр Мартин-Лефа. Псевдослучайные числа. Субъективное восприятие случайности.
4. Модификации модели, пренебрежимые с точки зрения интерпретации. Нуль-множества, пополнение вероятностного пространства. Непрерывные и дискретные модели, теорема Улама (частный случай, в предположении континуум-гипотезы). Роль аксиомы выбора, ее альтернативы, аксиома детерминированности. Борелевская сигма-алгебра. Теорема об изоморфизме.
5. Распределение вероятностей как частный вид модели. Случайные величины, вектора, функции. Интерпретация условия измеримости. Распределение как индуцированная мера, композиция измеримых отображений. Функция распределения как характеристика распределения и ее свойства. Моделирование случайных величин методом обращения функций распределения. Независимость в терминах распределений, произведение мер.
6. Альтернативный подход к аксиоматике на основе понятия среднего. Аксиома связи среднего и вероятности, эквивалентность подходу на основе понятия вероятности. Построение интеграла Лебега – приближение дискретными величинами. Теорема о замене меры в интеграле Лебега и ее вероятностная интерпретация. Мультипликативное свойство. Некоторые свойства математического

ожидания, неравенства Йенсена, Маркова-Чебышева.

7. Формулировка закона больших чисел и объективная интерпретация среднего. Субъективная интерпретация среднего. Отличие интеграла Лебега от интеграла Римана. Метод Монте-Карло как численный метод нахождения интеграла Лебега. Равномерно распределенные последовательности Соболя.
8. Ковариация, дисперсия и их свойства. Ковариационные матрицы. Линейные модели. Постановки задачи регрессии.
9. Абсолютно-непрерывные распределения, свойства плотности. Теорема Радона-Никодима. Гладкая замена переменных. Моделирование методом фон Неймана (методом элиминации). Дискретные распределения. Смеси распределений. Сингулярные распределения. Полная классификация распределений. Универсальность датчика случайных чисел.
10. Примеры часто употребляемых на практике распределений – биномиальное, геометрическое, экспоненциальное (модели и характеристики). Свойства многомерного нормального распределения. Моделирование гауссовского случайного вектора.
11. Совместные и маргинальные распределения. Согласованность семейства маргинальных распределений, конечный и бесконечный случай. Теорема Колмогорова и ее применения: система независимых случайных величин, гауссовские системы. Процессы с независимыми приращениями. Винеровский процесс.
12. Условные распределения (существование регулярного варианта), интегральные формулы полной вероятности. Случаи дискретного распределения известного случайного элемента и совместного абсолютно непрерывного распределения. Теорема Ионеску Тулча. Метод последовательного моделирования случайного вектора. Независимость последовательности случайных элементов.
13. Марковское свойство — интерпретация и две формализации. Уравнение Колмогорова-Чепмена. Функциональное представление для цепей Маркова (стохастическая динамическая система). Однородность по времени. Условие стационарности. Процесс Орнштейна-Уленбека.
14. Виды сходимости случайных величин: по вероятности, в среднем, почти наверное, по распределению — слабая и по вариации. Равномерная интегрируемость и предельный переход под знаком интеграла. Эргодичность марковских процессов.
15. Пуассоновское поле. Условия хаотичности, однородности и ординарности. Случай пуассоновского потока. Способы моделирования. Закон редких событий.
16. Характеристические, производящие функции и преобразование Лапласа как аппарат исследования сумм независимых случайных величин. Свойства некоторых распределений (нормального, биномиального, пуассоновского и т.д.). Доказательство закона больших чисел. Центральная предельная теорема. Устойчивые распределения.

А.2 Рекомендуемая литература

1. Ширяев А.Н. Вероятность. – в 2-х кн., 3-е изд. – М. МЦНМО, 2004
2. Тутубалин В.Н. Теория вероятностей и случайных процессов. - М., изд-во МГУ, 1992.
3. Соболев И.М. Численные методы Монте-Карло – М. Наука, 1973

А.3 Дополнительная литература

1. Колмогоров А.Н. Основные понятия теории вероятностей. Третье издание. – М. Фазис, 1998.
2. Уиттл П. Вероятность. - М. Наука, 1982.
3. Майстров Л.Е. Теория вероятностей: исторический очерк - М. Наука, 1967
4. Неве Ж. Математические основы теории вероятностей. - М. Мир, 1969.
5. Хеннекен П.А., Тортра А. Теория вероятностей и ее приложения - М. Наука, 1974.
6. Ермаков С.М. Метод Монте-Карло и смежные вопросы. - М. Наука, 1975.
7. Окстоби Дж. Мера и категория - М. Мир, 1974

Приложение В

Пример выполненного отчета

Отчет публикуется с разрешения автора.



Московский государственный университет имени М. В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра системного анализа

Отчёт по практикуму

«Стохастический анализ и моделирование»

Студент 415 группы
Д. М. Сотников

Руководитель практикума
А. Ю. Заночкин

Москва, 2020

Оглавление

1	Задание 1	33
1.1	Постановка задачи	33
1.2	Теоретическая часть	33
1.3	Результаты работы программы	35
2	Задание 2	37
2.1	Постановка задачи	37
2.2	Теоретическая часть	37
2.3	Результаты работы программы	39
3	Задание 3	40
3.1	Постановка задачи	40
3.2	Теоретическая часть	40
3.3	Результаты работы программы	43
4	Задание 4	46
4.1	Постановка задачи	46
4.2	Теоретическая часть	46
4.3	Результаты работы программы	47
5	Задание 5	49
5.1	Постановка задачи	49
5.2	Теоретическая часть	50
5.3	Результаты работы программы	51
6	Задание 6	53
6.1	Постановка задачи	53
6.2	Теоретическая часть	53

6.3	Результаты работы программы	55
7	Задание 7	56
7.1	Постановка задачи	56
7.2	Теоретическая часть	56
7.3	Результаты работы программы	58
8	Задание 8	60
8.1	Постановка задачи	60
8.2	Теоретическая часть	60
8.3	Результаты работы программы	61
9	Задание 9	62
9.1	Постановка задачи	62
9.2	Теоретическая часть	62
9.3	Результаты работы программы	65
10	Задание 10	66
10.1	Постановка задачи	66
10.2	Теоретическая часть	66
10.3	Результаты работы программы	67
11	Задание 11	70
11.1	Постановка задачи	70
11.2	Теоретическая часть	71
11.3	Результаты работы программы	73

1 Задание 1

1.1 Постановка задачи

1. Реализовать генератор схемы Бернулли с заданной вероятностью успеха p . Построить на его основе датчик биномиального распределения.
2. Реализовать генератор геометрического распределения. Проверить для него свойство отсутствия памяти.
3. Рассмотреть игру в орлянку — бесконечную последовательность независимых испытаний с бросанием правильной монеты. Выигрыш S_n определяется как сумма по n испытаниям значений 1 и -1 в зависимости от выпавшей стороны. Проиллюстрировать поведение $Y(i) = \frac{S_i}{\sqrt{n}}$ как функцию номера испытания $i = 1, \dots, n$ для отдельно взятой траектории. Дать теоретическую оценку $Y(n)$ при $n \rightarrow \infty$.

1.2 Теоретическая часть

Моделирование распределений. Для моделирования бернуллиевской случайной величины с вероятностью успеха p будем использовать следующий алгоритм:

- Моделируем равномерно распределенную на $[0, 1]$ случайную величину $U \sim \mathcal{U}(0, 1)$.
- Если $U < p$, то датчик возвращает $X = 1$, В противном случае $X = 0$.

Будем моделировать биномиальное распределение $\text{Bin}(n, p)$ как сумму n случайных величин, имеющих распределение Бернулли с параметром p .

Геометрическое распределение $\text{Geom}(p)$ можно моделировать с помощью схемы Бернулли с параметром p как число «неудач» до первого «успеха», однако этот метод требует работы с массивами случайной длины. Более эффективный подход дает следующее Пусть $\xi \sim \exp(\lambda)$, $\lambda > 0$. Тогда $\lfloor \xi \rfloor \sim \text{Geom}(p)$, где $p = 1 - e^{-\lambda}$.

Функция распределения ξ имеет вид $F_\xi(x) = 1 - e^{-\lambda x}$. Поэтому для $n \in \mathbb{N}_0$

$$(\lfloor \xi \rfloor = n) = (n \leq \xi < n + 1) = e^{-n\lambda}(1 - e^{-\lambda}) = (1 - p)^n p,$$

что и требовалось.

Таким образом, моделирование геометрического распределения с параметром p сводится к моделированию экспоненциального распределения с параметром $\lambda = -\ln(1 - p)$.

Экспоненциальное распределение легко моделируется при помощи метода обращения функции распределения:

$$F_\xi^{-1}(y) = -\frac{1}{\lambda} \ln(1 - y).$$

Пусть случайная величина ξ имеет непрерывную и строго возрастающую функцию распределения F_ξ . Тогда $F_\xi^{-1}(U) \sim \xi$, $U \sim \mathcal{U}(0, 1)$.

Непрерывность и строгое возрастание F_ξ гарантируют существование непрерывной и строго возрастающей обратной функции.

Найдем функцию распределения случайной величины $F_\xi^{-1}(U)$:

$$(F_\xi^{-1}(U) < x) = (U < F_\xi(x)) = F_\xi(x),$$

поэтому $F_\xi^{-1}(U) \sim \xi$.

Эту теорему легко обобщить и на более общий случай [1], однако в рамках данного практикума приведенной формулировки будет достаточно.

Отсутствие последействия. Будем говорить, что распределение дискретной случайной величины X обладает свойством отсутствия последействия, если $(X > t + \tau | X \geq \tau) = (X > t)$ для любых $t, \tau \in \mathbb{Z}$, $t > 0$.

У геометрического распределения $(X > t) = (1 - p)^{t+1}$, поэтому

$$(X > t + \tau | X \geq \tau) = \frac{(X > t + \tau, X \geq \tau)}{(X \geq \tau)} = \frac{(1 - p)^{t+\tau+1}}{(1 - p)^\tau} = (X > t).$$

Таким образом, геометрическое распределение обладает свойством отсутствия последействия. Более того, можно показать, что оно является единственным дискретным распределением, обладающим этим свойством. Поэтому проверка этого свойства для дискретного распределения позволяет определить, является ли оно геометрическим.

Игра в орлянку. Для анализа распределения $Y(i)$ из пункта 3. воспользуемся центральной предельной теоремой (далее ЦПТ): Пусть X_1, X_2, \dots — последовательность независимых случайных величин, имеющих конечное математическое ожидание μ и ненулевую дисперсию σ^2 . Тогда

$$\frac{S_n - \mu n}{\sigma \sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1) \text{ при } n \rightarrow \infty.$$

Ее доказательство можно найти в [9].

Поскольку при описанной игре в орлянку математическое ожидание равно 0, а дисперсия равна 1, распределение $Y(i)$ сходится к стандартному нормальному по ЦПТ.

1.3 Результаты работы программы

Проверка датчика биномиального распределения. На рис. 1 приведена гистограмма выборки, полученной с помощью построенного датчика. По ЦПТ такое распределение аппроксимируется нормальным с параметрами $\mu = np$, $\sigma^2 = np(1 - p)$.

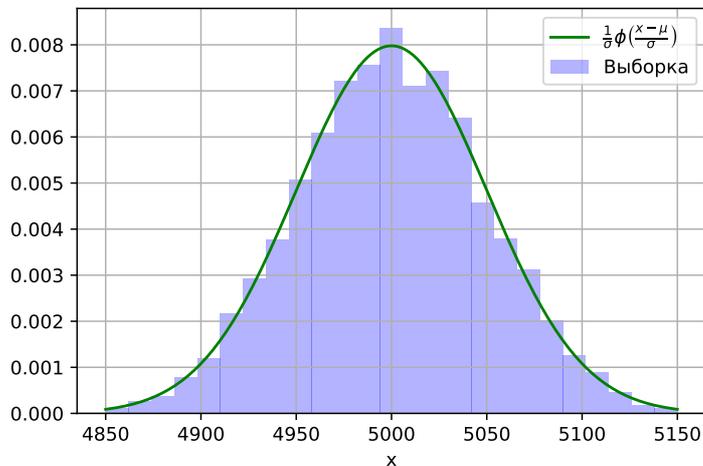


Рис. В.1: Моделирование биномиального распределения с $n = 10^4$, $p = 0.5$.

Проверка отсутствия памяти. Обозначим

$$\xi \sim \text{Geom}(p), \quad \eta = (\xi | \xi \geq \tau) - \tau.$$

В случае, когда распределение обладает свойством отсутствия последействия, ξ и η должны быть одинаково распределены. Рис. 2 показывает совпадение гистограмм и эмпирических функций распределения этих случайных величин при $p = 0.4$, $\tau = 3$.

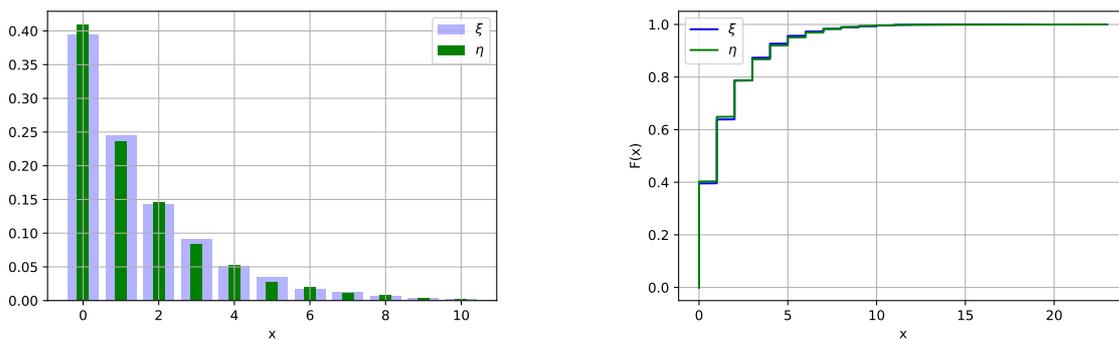


Рис. В.2: Сравнение гистограмм и выборочных функций распределения ξ и η .

Моделирование игры в орлянку. На рис. 3 представлены 10 траекторий случайной последовательности $Y(i) = \frac{S_i}{n}$. На рис. 4 построена гистограмма для выборки случайных величин $Y(n)$ при $n = 10^4$, подтверждающая теоретический результат.

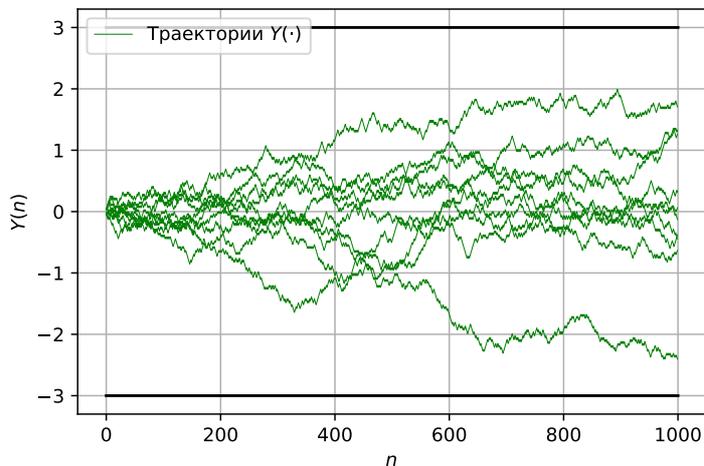


Рис. В.3: Траектории Y при игре в орлянку.

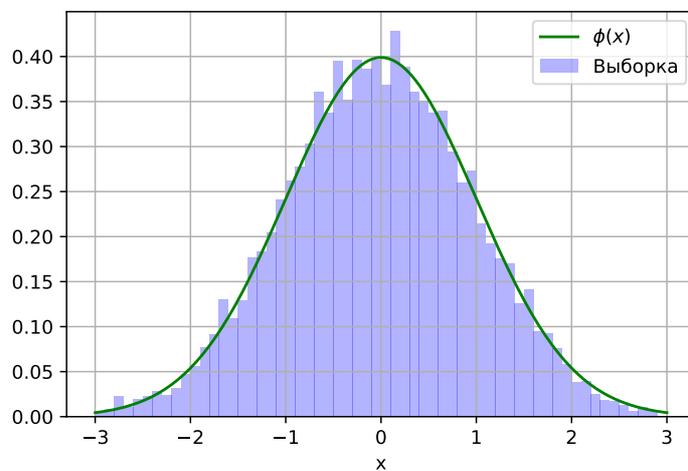


Рис. В.4: Выборка случайных величин, совпадающих по распределению с $Y(10^4)$.

2 Задание 2

2.1 Постановка задачи

1. Построить датчик сингулярного распределения, имеющий в качестве функции распределения канторову лестницу. С помощью критерия Колмогорова убедиться в корректности работы датчика.
2. Для канторовых случайных величин проверить свойство симметричности относительно 0.5 и самоподобия относительно деления на 3 с помощью критерия Смирнова.
3. Вычислить значение математического ожидания и дисперсии для данного распределения. Сравнить теоретические и эмпирические значения, проиллюстрировать сходимость.

2.2 Теоретическая часть

Моделирование сингулярного распределения. Для моделирования требуемого распределения воспользуемся тем, что носителем распределения является канторово множество, то есть числа, не имеющие единиц в троичной записи. Заметим также, что если k разрядов троичной записи случайной величины уже известны, то $(k + 1)$ -й разряд принимает одно из значений 0, 2 с одинаковой вероятностью. Это следует из самоподобия канторовой лестницы. Поэтому будем рассматривать схему Бернулли размера n , состоящую из разрядов числа в троичной системе исчисления (при «успехе» разряд равен 2) с таким n , чтобы ошибка в вычислении числа не превосходила ε .

Так как ошибка в худшем случае равна $\sum_{k=n+1}^{\infty} (2/3)^k = 3^{-n}$, достаточно взять $n \geq \log_3 \frac{1}{\varepsilon}$.

Критерий Колмогорова. Сформулируем теорему Колмогорова [4]. Пусть X_1, \dots, X_n — выборка независимых случайных величин, имеющих функцию распределения F . Построим по ней выборочную функцию распределения F_n . Тогда

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \xrightarrow{d} K,$$

где K — распределение Колмогорова.

Опишем алгоритм применения критерия Колмогорова. Будем проверять гипотезу

$$H_0: F(\cdot) = F_0(\cdot),$$

где $F_0(\cdot)$ — функция Кантора. Для этого вычислим

$$D_n = \sup_x |F_n(x) - F_0(x)|,$$

где $F_n(\cdot)$ — эмпирическая функция распределения, построенная по выборке размера n . Положим $\gamma = K^{-1}(1 - \alpha)$. Здесь α — уровень значимости, а $K^{-1}(x)$ — квантиль распределения Колмогорова порядка x .

Будем отвергать гипотезу в случае, когда

$$\sqrt{n}D_n \geq \gamma.$$

Положим $\alpha = 0.05$. Тогда из таблицы распределения Колмогорова $\gamma = K^{-1}(0.95) = 1.36$.

Критерий Смирнова. Для того, чтобы по двум выборкам проверить, имеют ли они одинаковое распределение, используется критерий Смирнова.

Рассматриваются $X = (x_1, \dots, x_n)$, $Y = (y_1, \dots, y_m)$ — выборки, элементы которых имеют функции распределения F и G соответственно, а F_n, G_m — построенные по ним эмпирические функции распределения. Обозначим

$$D_{mn} = \sup_{x \in \mathbb{R}} |F_n(x) - G_m(x)|.$$

Теорема Смирнова [5] утверждает, что

$$\mathbb{P} \left(\sqrt{\frac{mn}{m+n}} D_{mn} < x \right) \rightarrow K(x).$$

Будем отвергать гипотезу

$$H_0: F(\cdot) = G(\cdot),$$

если $\sqrt{\frac{mn}{m+n}} D_{mn} > \gamma$, $\gamma = K^{-1}(1 - \alpha)$.

Как и в прошлом пункте, $\alpha = 0.05$, $\gamma = 1.36$.

Для проверки симметричности будем проверять гипотезу

$$H_0: X \stackrel{d}{=} 1 - X.$$

с помощью критерия Смирнова, моделируя выборки с помощью построенного датчика.

Проверка самоподобия сводится к проверке гипотезы

$$H_0: \frac{Y}{3} \stackrel{d}{=} \left(Y \mid Y \leq \frac{1}{3} \right).$$

Моменты сингулярного распределения. В силу симметричности распределения относительно $\frac{1}{2}$, $\mathbb{E}X = \frac{1}{2}$. Условия симметричности и самоподобия можно переписать в виде

$$F(x) = 1 - F(1 - x), \quad F\left(\frac{x}{3}\right) = \frac{1}{2}F(x), \quad x \in [0, 1].$$

Вычислим второй момент, пользуясь этими свойствами:

$$\mathbb{E}X^2 = \int_0^1 x^2 dF(x) = \int_0^{\frac{1}{3}} x^2 dF(x) + \int_{\frac{2}{3}}^1 x^2 dF(x).$$

После замен переменных, переводящих множество интегрирования в отрезок $[0, 1]$, используя свойства симметричности и самоподобия, получим

$$\mathbb{E}X^2 = \frac{1}{3} + \frac{1}{9}\mathbb{E}X^2, \quad \mathbb{E}X^2 = \frac{3}{8},$$

откуда легко найти $X = \frac{1}{8}$.

Оценим по выборке математическое ожидание и дисперсию, используя статистики

$$\bar{x} = \frac{1}{n} \sum_{k=1}^n x_k, \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2.$$

2.3 Результаты работы программы

Проверка датчика. Рис. 5 демонстрирует работу построенного датчика сингулярного распределения. Размер выборки $n = 1000$.

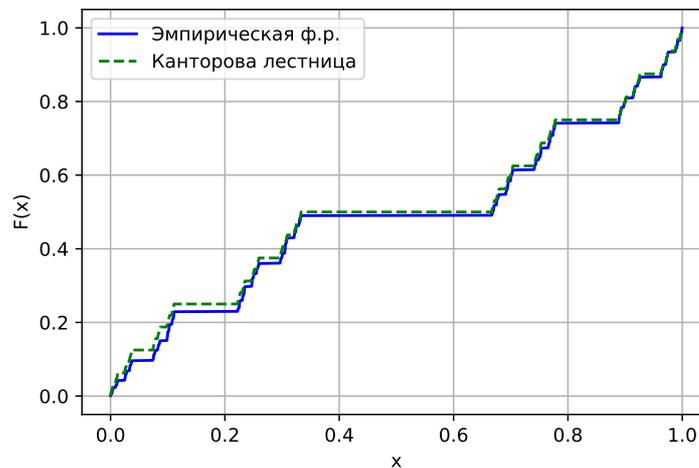


Рис. В.5: Работа датчика сингулярного распределения.

Проверка гипотез. Проверим построенные датчик с помощью критерия Колмогорова с уровнем значимости $\alpha = 0.05$ и размером выборки $n = 1000$. При моделировании проверки гипотезы $m = 1000$ раз, нулевая гипотеза отклонялась в 4.5% случаев.

При проверке симметричности и самоподобия с помощью критерия Смирнова с теми же параметрами, нулевая гипотеза отклонялась ровно в 5% случаев для обеих гипотез.

Оценка моментов. На рис. 6 показана сходимость выборочных математического ожидания и дисперсии к найденным теоретически показателям.

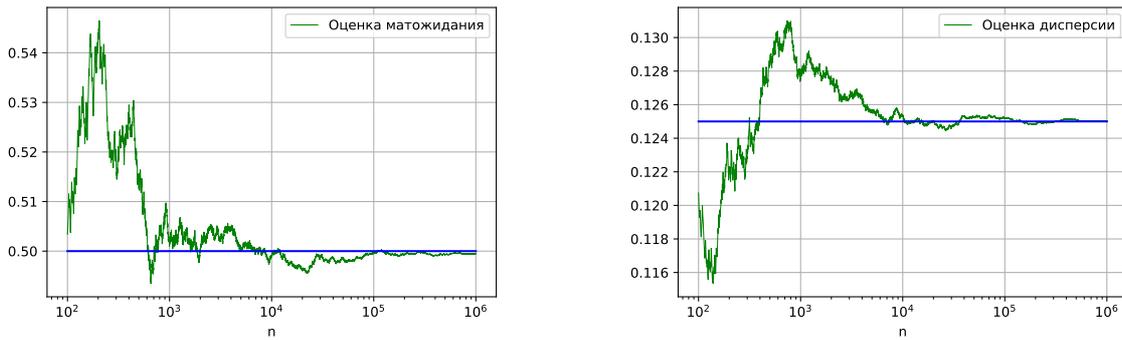


Рис. В.6: Оценка моментов сингулярного распределения.

3 Задание 3

3.1 Постановка задачи

1. Построить датчик экспоненциального распределения. Проверить для данного распределения свойство отсутствия памяти. Пусть X_1, \dots, X_n независимы и имеют экспоненциальное распределение с параметрами $\lambda_1, \dots, \lambda_n$. Найти распределение случайной величины $Y = \min(X_1, \dots, X_n)$.
2. На основе датчика экспоненциального распределения построить датчик пуассоновского распределения.
3. Построить датчик пуассоновского распределения как предел биномиального. С помощью критерия хи-квадрат убедиться, что получен датчик распределения Пуассона.
4. Построить датчик стандартного нормального распределения методом моделирования случайных величин с переходом в полярные координаты. Проверить при помощи t-критерия Стьюдента равенство матожиданий, а при помощи критерия Фишера равенство дисперсий.

3.2 Теоретическая часть

Экспоненциальное распределение. Датчик экспоненциального распределения уже был построен в задании 1. Свойство отсутствия последействия формулируется и доказывается в этом случае аналогично тому, как это было сделано для геометрического распределения.

Найдем теперь распределение $Y = \min(X_1, \dots, X_n)$:

$$\begin{aligned}
 F_Y(x) &= \mathbb{P}(\min(X_1, \dots, X_n) < x) = 1 - \mathbb{P}(X_1 \geq x, \dots, X_n \geq x) = \\
 &= 1 - \prod_{i=1}^n (1 - F_{X_i}(x)) = 1 - \exp\left(-\sum_{i=1}^n \lambda_i x\right),
 \end{aligned}$$

откуда $Y \sim \exp(\sum_{k=1}^n \lambda_k)$.

Моделирование распределения Пуассона. Рассмотрим пуассоновский процесс $X(t)$ с интенсивностью λ . Так как пуассоновский процесс является марковским, время между его скачками имеет экспоненциальное распределение с параметром λ . При этом $X(1)$ имеет распределение Пуассона с параметром λ . Таким образом, при моделировании $\xi_k \sim \exp(\lambda)$ до тех пор, пока

$$\sum_{k=1}^{\eta} \xi_k > 1,$$

величина $\eta - 1$, равная числу скачков до момента $t = 1$, будет иметь требуемое пуассоновское распределение.

Другой подход к построению датчика дает теорема Пуассона (доказательство в [9]). Рассмотрим схему серий испытаний Бернулли с вероятностями успеха p_n . Тогда, если

$$p_n \xrightarrow{n \rightarrow \infty} 0, \quad np_n \xrightarrow{n \rightarrow \infty} \lambda > 0,$$

число успехов μ_n в схеме Бернулли сходится по распределению к пуассоновскому распределению с параметром λ .

При этом верна оценка

$$\left| \mathbb{P}(\nu_n \in A) - \sum_{k \in A} e^{-\lambda} \frac{\lambda^k}{k!} \right| \leq np^2 = \lambda p.$$

Потребуем, чтобы вероятности отличались не более, чем на ε . Тогда биномиальное распределение с параметрами $p = \frac{\varepsilon}{\lambda}$, $n = \frac{\lambda}{p}$ аппроксимирует требуемое распределение пуассона с заданной точностью.

Критерий χ^2 Пирсона. Критерий χ^2 используется для проверки гипотезы

$$H_0: F = F_0,$$

где F_0 — заданная функция распределения. Разобьем на непересекающиеся множества $\Delta_1, \dots, \Delta_m$. Обозначим $p_i = (\Delta_i | H_0)$, N_i — число элементов выборки X_1, \dots, X_n , попавших в Δ_i . Тогда статистика

$$Q = \sum_{i=1}^m \frac{(N_i - np_i)^2}{np_i}$$

имеет¹ распределение $\chi^2(m-1)$.

Будем отклонять H_0 , если $Q \geq \gamma$, где $\gamma = \chi_{m-1}^2(1-\alpha)$, а $\chi_{m-1}^2(x)$ — квантиль порядка x распределения хи-квадрат с $m-1$ степенью свободы. Построенный таким образом критерий имеет уровень согласия α .

¹Формулировки всех утверждений, относящихся к проверке гипотез, приведены в [2]

Моделирование нормального распределения. Ясно, что достаточно построить датчик X стандартного нормального распределения. Нормальное распределение с параметрами μ, σ^2 получается из него преобразованиями сдвига и масштаба $\mu + \sigma X$.

Стандартное нормальное распределение будем генерировать полярным методом: рассмотрим двумерную случайную величину (X_1, X_2) , координаты которой имеют стандартное нормальное распределение. Найдем распределения полярных координат. Очевидно, что в силу симметрии угол ϕ имеет равномерное распределение на отрезке $[0, 2\pi]$. Квадрат радиуса по определению имеет распределение χ^2 с двумя степенями свободы, которое совпадает с $\exp\left(\frac{1}{2}\right)$. Таким образом, полярные координаты моделируются доступными датчиками распределений

$$r^2 \sim \exp\left(\frac{1}{2}\right), \quad \phi \sim \mathcal{U}(0, 2\pi),$$

откуда легко перейти к декартовым координатам

$$x = r \cos \phi \sim \mathcal{N}(0, 1), \quad y = r \sin \phi \sim \mathcal{N}(0, 1).$$

Критерий Стьюдента. Рассматривается выборка X_1, \dots, X_n из нормального распределения с неизвестными параметрами. Критерий Стьюдента позволяет проверить гипотезу

$$H_0: \mathbb{E}X = \mu.$$

Для оценки стандартного отклонения используется оценка

$$\sigma' = \sqrt{\frac{1}{n-1} \sum_i (X_i - \bar{X})^2}.$$

Тогда статистика

$$U = \sqrt{n} \frac{\bar{X} - \mu}{\sigma'}.$$

имеет распределение Стьюдента с $n - 1$ степенью свободы (при условии истинности H_0).

Для построения критерия с уровнем значимости α будем отклонять H_0 , когда $|U| \geq \gamma$, где $\gamma = T_{n-1}^{-1}\left(1 - \frac{\alpha}{2}\right)$.

Критерий Фишера. Критерий Фишера используется для сравнения дисперсий двух нормальных выборок с неизвестными параметрами. Точнее, будем проверять гипотезу

$$H_0: \sigma_1 = \sigma_2,$$

где σ_1, σ_2 — стандартные отклонения нормальных распределений, из которых взяты первая и вторая выборки соответственно.

Рассмотрим выборки $X = (X_1, \dots, X_n)$, $Y = (Y_1, \dots, Y_m)$. Для них статистика

$$V = \frac{m-1}{n-1} \frac{\sum_i (X_i - \bar{X})^2}{\sum_i (Y_i - \bar{Y})^2}$$

имеет распределение Фишера с параметрами $n - 1$ и $m - 1$.

Будем отклонять H_0 , если $V \leq \gamma_1$ или $V \geq \gamma_2$. Здесь

$$\gamma_1 = F_{n-1, m-1}^{-1}\left(\frac{\alpha}{2}\right), \quad \gamma_2 = F_{n-1, m-1}^{-1}\left(1 - \frac{\alpha}{2}\right),$$

а $F_{n-1, m-1}^{-1}(x)$ — квантиль F-распределения с параметрами $n - 1$ и $m - 1$ порядка x .

3.3 Результаты работы программы

Отсутствие последствия экспоненциального распределения. Как и в задании 1,

$$\xi \sim \text{Exp}(\lambda), \quad \eta = (\xi | \xi \geq \tau) - \tau.$$

На рис. 7 показано совпадение гистограмм и выборочных функций распределения при $\lambda = 0.5$, $\tau = 3$.

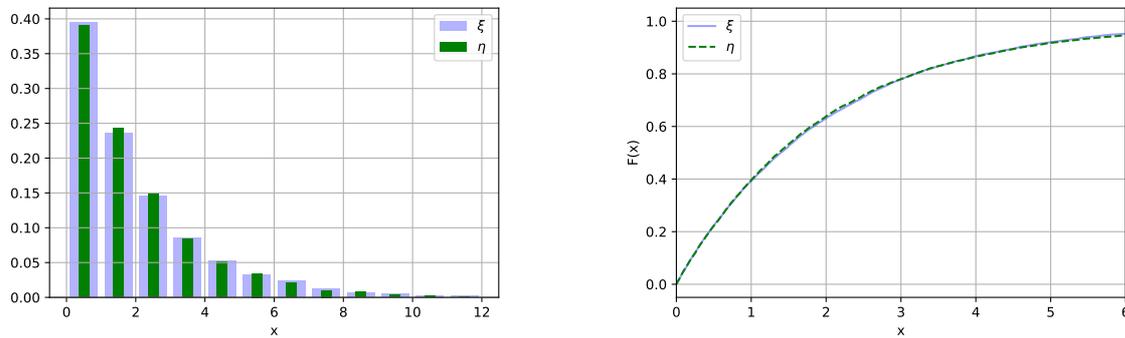


Рис. В.7: Сравнение гистограмм и выборочных функций распределения ξ и η .

Распределение $\min(X_1, \dots, X_n)$. На рис. 8 показано эмпирическое распределение минимума трех случайных величин X_1, X_2, X_3 , имеющих экспоненциальные распределения с параметрами $\lambda_1 = 2$, $\lambda_2 = 1$, $\lambda_3 = 4$. Зеленая линия на графике — плотность экспоненциального распределения с параметром $\lambda = 7$.

Критерий Пирсона. Во всех статистических критериях в этом задании будем брать уровень значимости $\alpha = 0.05$, размер выборки $n = 1000$. Процесс проверки гипотезы будет моделироваться $m = 1000$ раз. Для применения критерия Пирсона для распределения Пуассона с $\lambda = 4$, разобьем множество целых неотрицательных чисел на подмножества

$$\Delta_1 = \{0, 1\}, \Delta_2 = \{2\}, \Delta_3 = \{3\}, \Delta_4 = \{4\}, \Delta_5 = \{5\}, \Delta_6 = \{6\}, \Delta_7 = \{k \geq 7\}.$$

Вычислим вероятности $p_i = \mathbb{P}(\Delta_i)$ попадания в каждое множество

$$p_1 \approx 0.0916, \quad p_2 \approx 0.1465, \quad p_3 \approx 0.1954, \quad p_4 \approx 0.1954, \quad p_5 \approx 0.1563, \quad p_6 \approx 0.1042, \quad p_7 \approx 0.1107.$$

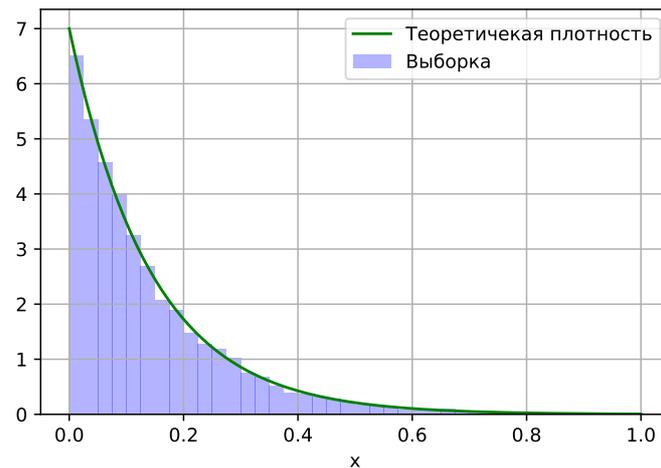


Рис. В.8: Распределение $Y = \min(X_1, \dots, X_n)$.

Множества Δ_i выбраны таким образом, чтобы теоретические вероятности попадания в них имели один порядок.

При моделировании критерия Пирсона нулевая гипотеза была отклонена в 4.3% случаев.

Работа полярного метода. Результат работы датчика нормального распределения, использующего полярный метод, показан на гистограмме на рис. 9.

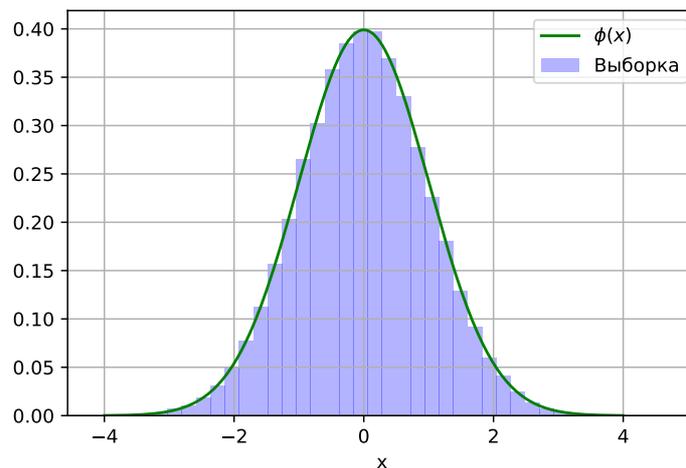


Рис. В.9: Моделирование стандартного нормального распределения полярным методом.

Критерий Стьюдента. С помощью датчика нормального распределения с $\mu = 1$, $\sigma^2 = 4$ была сгенерирована выборка. Гипотеза

$$H_0 : \mu = 1$$

отклонялась критерием Стьюдента в 3.1% случаев.

Критерий Фишера. Для тех же параметров $\mu = 1$, $\sigma^2 = 4$ гипотеза

$$H_0 : \sigma^2 = 4$$

отклонялась критерием Фишера в 5.6% случаев.

4 Задание 4

4.1 Постановка задачи

1. Построить датчик распределения Коши.
2. На основе датчика распределения Коши с помощью метода фон Неймана построить датчик стандартного нормального распределения. Убедиться в корректности построенного датчика и обосновать наблюдаемую линейную зависимость.
3. Сравнить скорость моделирования стандартного нормального распределения в заданиях 3 и 4.

4.2 Теоретическая часть

Моделирование распределения Коши. Функция распределения случайной величины, имеющей распределение Коши с параметром сдвига x_0 и параметром масштаба γ , имеет вид

$$F(x) = \frac{1}{\pi} \arctan\left(\frac{x - x_0}{\gamma}\right) + \frac{1}{2}.$$

Обратная к ней

$$F^{-1}(y) = x_0 + \gamma \tan\left(\pi\left(y - \frac{1}{2}\right)\right).$$

Будем использовать ее для моделирования распределения Коши:

$$\xi = F^{-1}(U) \sim \mathcal{C}(x_0, \gamma), \quad U \sim \mathcal{U}(0, 1).$$

Метод элиминации фон Неймана. Пусть имеется датчик распределения с плотностью $g(\cdot)$ (в этой задаче $g(x) = \frac{\gamma}{\pi[(x-x_0)^2 + \gamma^2]}$), и требуется смоделировать случайную величину с плотностью $f(\cdot)$ (в нашем случае $f(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$), причем

$$\sup_{x \in \mathbb{R}} \frac{f(x)}{g(x)} \leq k.$$

Будем использовать следующий алгоритм:

1. Моделируем $X \sim g$ и равномерно распределенную на единичном отрезке случайную величину U .
2. Если $\frac{f(X)}{g(X)} \geq kU$, то $Y = X$, и процесс завершается, в противном случае нужно вернуться у шагу 1.

Докажем, что полученная таким образом случайная величина Y имеет требуемое распределение с плотностью f . Обозначим $A = \left\{ \omega : \frac{f(X)}{g(X)} \geq kU \right\}$.

$$(A) = \int_0^1 \frac{f(x)}{kg(x)} g(x) du dx = \frac{1}{k} \int_f (x) dx = \frac{1}{k}.$$

Условная плотность (X, U) при условии A имеет вид $p(x, u|A) = kg(x)$, если A выполнено.

Тогда для произвольного борелевского B

$$(X \in B|A) = \int_B \frac{f(x)}{kg(x)} kg(x) du dx = \int_B f(x) dx,$$

что и требовалось доказать.

Нетрудно показать, что наилучшее приближение стандартного нормального распределения распределением Коши имеет параметры $x_0 = 0, \gamma = 1$. При этом

$$k = \sup_x \frac{f(x)}{g(x)} = \sqrt{\frac{2\pi}{e}}.$$

Обоснование линейной зависимости в `normplot`. Для стандартного нормального распределения точки лежат на прямой по построению: ось ординат преобразуется так, чтобы функция нормального распределения была линейной. Таким образом, если $\xi \sim \mathcal{N}(0, 1)$, то $F(x) = \mathbb{P}(\xi < x)$ линейна на таком графике. Произвольную нормальную случайную величину можно представить в виде $\eta = \sigma\xi + \mu$, и для нее

$$\mathbb{P}(\eta < x) = \mathbb{P}\left(\xi < \frac{x - \mu}{\sigma}\right)$$

также будет линейной на этом графике. При этом происходит сдвиг на μ по оси абсцисс, а угловой коэффициент умножается на $\frac{1}{\sigma}$.

Сравнение скорости моделирования нормального распределения. Для сравнения скорости работы построенных датчиков, создадим логарифмическую сетку для размеров выборки размера n . Для выборки каждого размера будем запускать датчик m раз, а затем в качестве оценки времени работы брать медиану измерений как устойчивую к большим выбросам оценку среднего.

4.3 Результаты работы программы

Датчик распределения Коши На рис. 10 показана работа датчика, моделирующего стандартное распределение Коши.

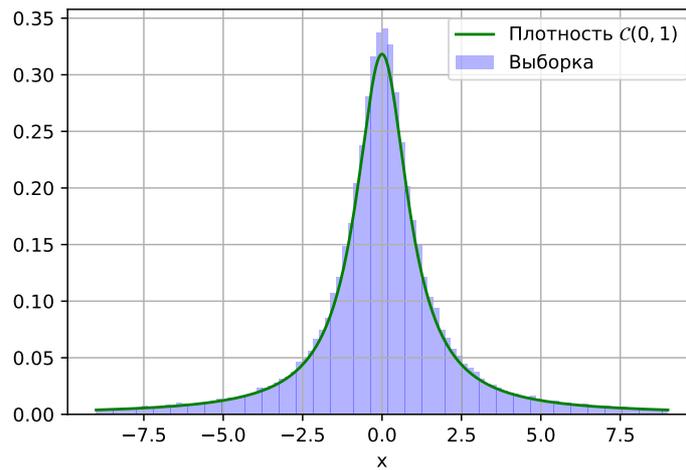
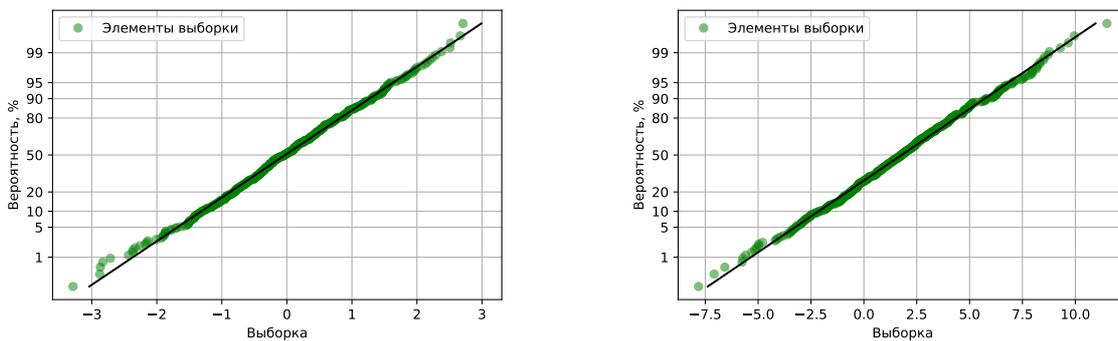


Рис. В.10: Моделирование стандартного распределения Коши.

Рис. В.11: normplot для $\mathcal{N}(0, 1)$ и $\mathcal{N}(2, 9)$.

Графики normplot. На рис. 11 изображены графики normplot для стандартного нормального распределения (слева) и для нормального распределения с параметрами $\mu = 2$, $\sigma^2 = 3$. Рис. 12 показывает, что линейная зависимость пропадает для выборки из распределения, не являющегося нормальным. Также из графика видно, что распределение Коши имеет значительно более тяжелые хвосты, нежели нормальное.

Сравнение времени работы датчиков. На рис. 13 показана зависимость времени работы датчиков от размера выборки. Сравняются датчики, использующие полярный метод и метод элиминации фон Неймана, а также стандартный датчик, реализованный в библиотеке numru. Каждый датчик запускается $m = 100$ раз, среднее время оценивается с помощью медианы. Таким образом редкие большие выбросы, возможные при работе компьютера, не оказывают сильного влияния на среднюю величину.

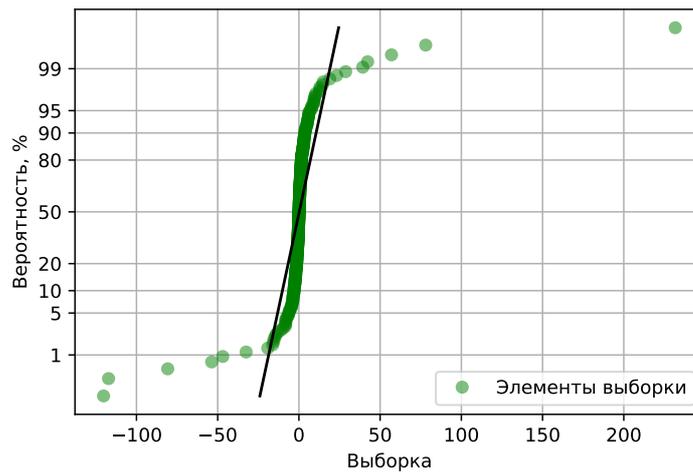


Рис. В.12: normplot для стандартного распределения Коши.

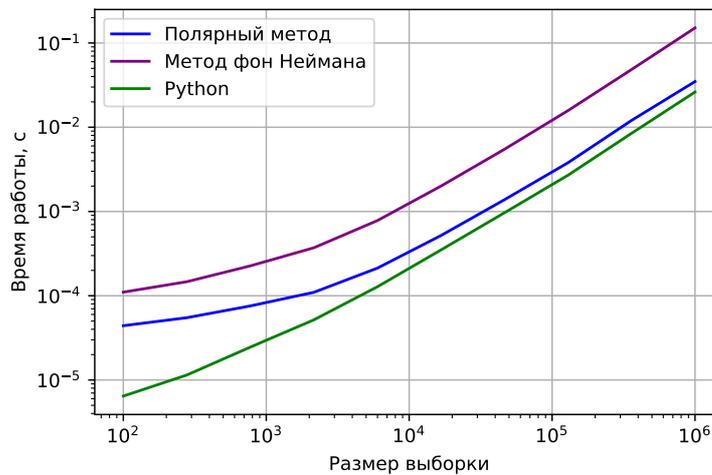


Рис. В.13: Сравнение времени работы датчиков.

5 Задание 5

5.1 Постановка задачи

1. Пусть $X_i \sim \mathcal{N}(\mu, \sigma^2)$. Убедиться эмпирически в справедливости ЗБЧ и ЦПТ: исследовать поведение суммы S_n и эмпирического распределения величины

$$\sqrt{n} \left(\frac{S_n}{n} - \mu \right).$$

2. Считая μ и σ неизвестными, для пункта 1 построить доверительные интервалы для среднего и дисперсии.
3. Пусть $X_i \sim \mathcal{C}(a, b)$ имеет распределение Коши со сдвигом a и масштабом b . Проверить эмпирически, как ведут себя суммы $\frac{S_n}{n}$. Результат объяснить, а также найти распределения данных сумм.

5.2 Теоретическая часть

ЦПТ уже была сформулирована как теорема 2. Приведем теперь формулировку усиленного закона больших чисел (далее УЗБЧ) в форме Колмогорова. Его доказательство можно найти, например, в [9]. Пусть X_1, X_2, \dots — последовательность независимых одинаково распределенных случайных величин, имеющих конечные первые моменты. Тогда

$$\frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{\text{п.н.}} X_1.$$

По ЗБЧ величина S_n должна вести себя асимптотически как nX_1 , а случайная величина $\sqrt{n} \left(\frac{S_n}{n} - \mu \right)$ сходится по распределению к $\mathcal{N}(0, \sigma^2)$.

Доверительный интервал для математического ожидания. Так как статистика

$$U = \sqrt{n} \frac{\bar{X} - \mu}{\sigma'}, \quad \sigma' = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1}}$$

имеет [2] распределение Стьюдента с $n-1$ степенью свободы, можно построить доверительный интервал с уровнем доверия α следующим образом:

$$\mathbb{P}(\gamma_1 < U < \gamma_2) = \mathbb{P}\left(\bar{X} - \frac{\sigma' \gamma_2}{\sqrt{n}} < \mu < \bar{X} - \frac{\sigma' \gamma_1}{\sqrt{n}}\right) = \alpha.$$

Значения γ_1 и γ_2 выберем так, чтобы интервал был симметричным:

$$\gamma_1 = T_{n-1}^{-1}\left(\frac{1-\alpha}{2}\right), \quad \gamma_2 = T_{n-1}^{-1}\left(\frac{1+\alpha}{2}\right).$$

Итак, доверительный интервал для параметра μ имеет вид

$$\left(\bar{X} - \frac{\sigma' \gamma_2}{\sqrt{n}}, \bar{X} - \frac{\sigma' \gamma_1}{\sqrt{n}}\right).$$

Доверительный интервал для дисперсии. Для оценки дисперсии будем использовать статистику

$$V = \frac{\sum_i (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2(n-1).$$

$$\mathbb{P}(\gamma_1 < V < \gamma_2) = \mathbb{P}\left(\frac{\sum_i (X_i - \bar{X})^2}{\gamma_2} < \sigma^2 < \frac{\sum_i (X_i - \bar{X})^2}{\gamma_1}\right) = \alpha.$$

Для построения симметричного интервала

$$\gamma_1 = \chi_{n-1}^{-1}\left(\frac{1-\alpha}{2}\right), \quad \gamma_2 = \chi_{n-1}^{-1}\left(\frac{1+\alpha}{2}\right).$$

Доверительный для σ^2 с уровнем доверия α имеет вид

$$\left(\frac{\sum_i (X_i - \bar{X})^2}{\gamma_2}, \frac{\sum_i (X_i - \bar{X})^2}{\gamma_1}\right).$$

Поведение сумм распределения Коши. Найдем распределение $\frac{S_n}{n}$ с помощью характеристической функции распределения Коши:

$$\varphi(t) = \mathbb{E}e^{it\xi} = \frac{1}{\pi} \int \frac{be^{itx}}{(x-a)^2 + b^2} dx = \frac{e^{ita}}{\pi} \int \frac{e^{itbx}}{x^2 + 1} dx = e^{ita - |t|b}.$$

Тогда характеристическая функция $\frac{S_n}{n}$ имеет вид

$$\varphi_{\frac{S_n}{n}}(t) = \left[\varphi\left(\frac{t}{n}\right) \right]^n = \varphi(t),$$

поэтому $\frac{S_n}{n} \sim \mathcal{C}(a, b)$.

5.3 Результаты работы программы

Иллюстрация ЗБЧ. Рис. 14 демонстрирует сходимость нормированных сумм к среднему при проведении дополнительных испытаний.

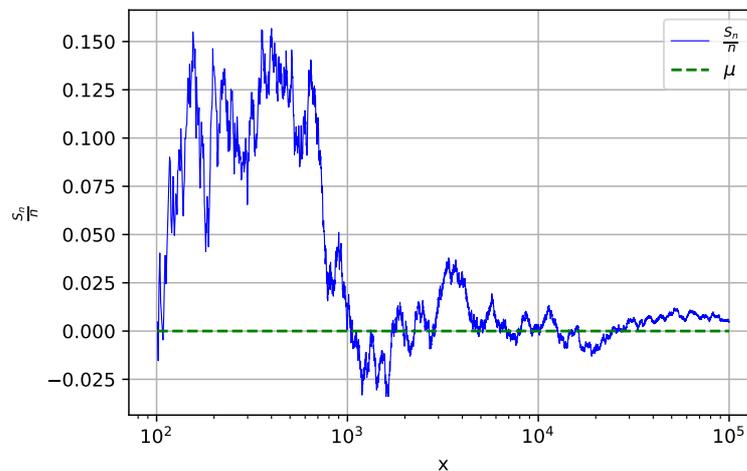


Рис. В.14: Поведение нормированных сумм.

Иллюстрация ЦПТ. На рис. 15 показана гистограмма выборки случайной величины $\sqrt{n} \left(\frac{S_n}{n} - \mu \right)$, а также плотность теоретического предела данной случайной последовательности.

Доверительные интервалы. На рис. 16 изображены доверительные интервалы с уровнем доверия $\alpha = 0.95$ для параметров μ , σ^2 , построенные в соответствии с теоретической частью. При построении по выборке размера $n = 1000$ доверительных интервалов $m = 1000$ раз, истинный параметр оказывался в интервале в 95.4% случаев для математического ожидания и в 95.3% случаев для дисперсии.

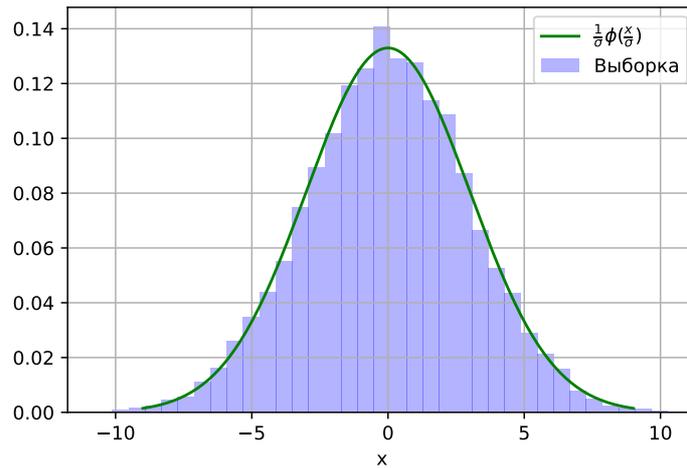


Рис. В.15: Выборка из $\sqrt{n} \left(\frac{S_n}{n} - \mu \right)$ при $n = 10^4$.

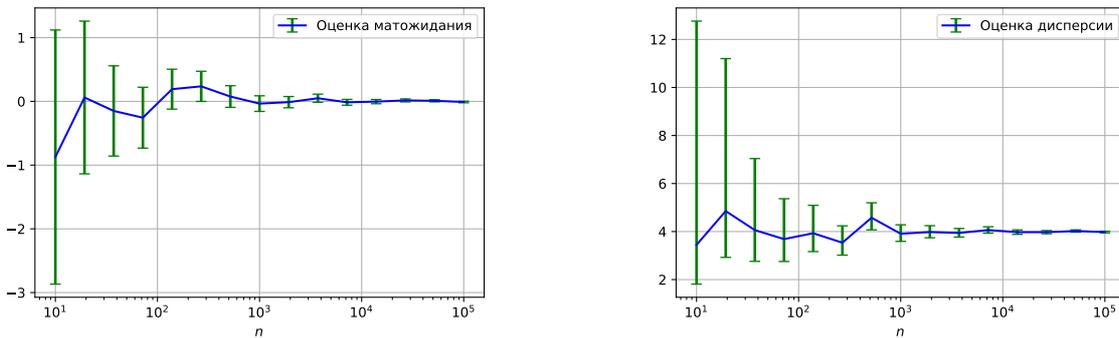


Рис. В.16: Доверительные интервалы для μ и σ^2 . Истинные значения $\mu = 0$, $\sigma = 2$.

Поведение сумм случайных величин, имеющих распределение Коши. На рис. 17 изображено поведение нормированной суммы, каждое слагаемое которой имеет распределение Коши с параметром сдвига 1 и параметром масштаба, равным 2. Из-за того, что отдельные элементы вносят значительный вклад в общую сумму, сходимость отсутствует, и распределение нормированной суммы является распределением Коши с теми же параметрами.

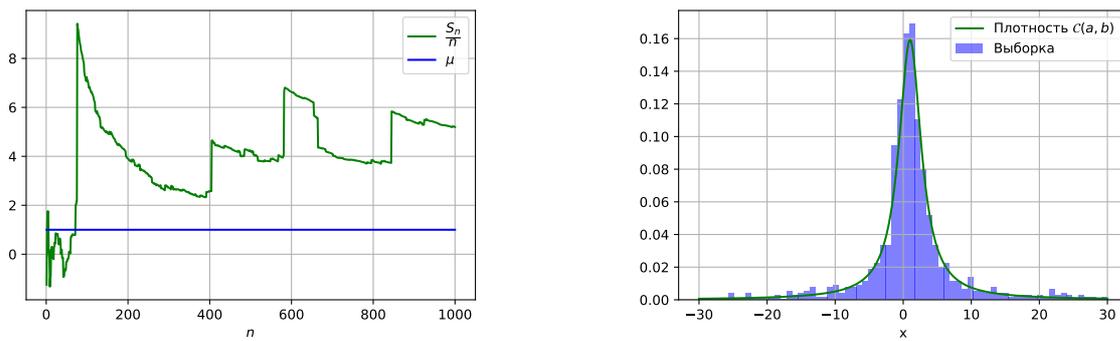


Рис. В.17: «ЗБЧ» для распределения Коши и распределение $\frac{S_n}{n}$ при $n = 10^4$.

6 Задание 6

6.1 Постановка задачи

1. Посчитать интеграл

$$J = \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \dots \int_{-\infty}^{+\infty} \frac{e^{-\left(x_1^2 + \dots + x_{10}^2 + \frac{1}{27x_1^2 \dots x_{10}^2}\right)}}{x_1^2 \dots x_{10}^2} dx_1 dx_2 \dots dx_{10}$$

- методом Монте–Карло;
- методом квадратур, сводя задачу к вычислению собственного интеграла Римана.

2. Для каждого случая оценить точность вычислений.

6.2 Теоретическая часть

Метод Монте–Карло. Отметим, что случайный вектор $X = (X_1, \dots, X_{10})$, у которого каждая компонента распределена нормально с $\mu = 0$, $\sigma^2 = \frac{1}{2}$, имеет плотность

$$p(x) = \frac{1}{\pi^5} e^{-(x_1^2 + \dots + x_{10}^2)}, \quad x \in \mathbb{R}^{10}.$$

Перепишем интеграл в виде

$$J = \pi^5 \int_{\mathbb{R}^{10}} y e^{-\frac{y}{128}} p(x) dx, \quad y = \left(\prod_{k=1}^{10} x_k \right)^{-2}$$

и заметим, что он равен математическому ожиданию $\mathbb{E}f(Y)$, где $f(y) = \pi^5 y e^{-\frac{y}{128}}$, $Y = \left(\prod_{k=1}^{10} X_k \right)^{-2}$, а $X_k \sim \mathcal{N}\left(0, \frac{1}{2}\right)$ и независимы.

По усиленному закону больших чисел

$$\frac{1}{n} \sum_{k=1}^n f(Y_k) \xrightarrow{\text{п.н.}} \mathbb{E}f(Y) = J.$$

Оценка точности. Для метода Монте-Карло найдем ошибку ε , вероятность превзойти которую при моделировании выборки размера n не превосходит α . Воспользуемся для этого неравенством Чебышёва:

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{k=1}^n f(Y_k) - \mathbb{E}f(Y)\right| \geq \varepsilon\right) \leq \frac{[f(Y)]}{n\varepsilon^2} = \alpha.$$

Дисперсию $f(Y)$ будем аппроксимировать выборочной дисперсией $\frac{1}{n}\sum_i (f(Y_i) - \overline{f(Y)})^2$, которая является состоятельной оценкой.

Метод квадратур. Сделаем в интеграле замену

$$x_k = \tan(y_k), \quad dx_k = \frac{dy_k}{\cos^2(y_k)}, \quad k = 1, \dots, 10.$$

После замены и некоторых упрощений интеграл примет вид

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \dots \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \exp\left\{-\sum_k \tan^2(y_k) - \frac{1}{128} \prod_k \cot^2(y_k)\right\} \prod_k \frac{1}{\sin^2(y_k)} dy_1 \dots dy_{10}$$

Обозначим подынтегральную функцию $f(y_1, \dots, y_{10})$ и заметим, что она симметрична и четна по всем аргументам.

Построим равномерную сетку размера m на каждом отрезке $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$:

$$x_i = -\frac{\pi}{2} + h(i + 0.5), \quad i = 0, \dots, m - 1,$$

В силу четности f можно вычислить интеграл на $\left[0, \frac{\pi}{2}\right]$ по каждому измерению, и затем умножить результат на 2^{10} .

$h = \frac{\pi}{m}$ — шаг сетки. Элементы сетки — центры отрезков, полученных при разбиении $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]$ на m равных частей. Разобьем $\left[-\frac{\pi}{2}, \frac{\pi}{2}\right]^{10}$ на гиперкубики с центрами

$$(x_{i_1}, \dots, x_{i_{10}}), \quad i_k = 0, \dots, m - 1, \quad k = 1, \dots, 10$$

и мерой h^{10} и воспользуемся методом прямоугольников

$$J \approx h^{10} \sum_{i_1=0}^{m-1} \dots \sum_{i_{10}=0}^{m-1} f(x_{i_1}, \dots, x_{i_{10}})$$

Для вычисления этой суммы потребуется m^{10} вычислений функции f , что уже при $m = 5$ потребует около 10^7 итераций. Для оптимизации вычислений воспользуемся симметричностью f :

$$J \approx \sum_{0 \leq i_1 \leq \dots \leq i_{10} \leq m-1} \frac{10!}{n_0! \dots n_{m-1}!} f(x_{i_1}, \dots, x_{i_{10}}),$$

где n_i обозначается число индексов, равных i на данной итерации. Число слагаемых при этом равно $\binom{m+9}{10}$.

Оценка точности. Для оценки погрешности будем использовать разложение в ряд Тейлора с остаточным членом в форме Лагранжа функции f :

$$f(y) = f(y_0) + \langle \nabla f(\xi), \Delta y \rangle,$$

где y_0 центр гиперкубика Δ , а $\xi \in \Delta$. Интегрируя обе части по Δ , получим, что ошибка интегрирования по каждому кубу равна

$$\int_{\Delta} \langle \nabla f(\xi), \Delta y \rangle dy \leq \sup_{\Delta} \|\nabla f(x)\| \|\Delta y\| \mu(\Delta) \leq \frac{\sqrt{n}}{2} h^{n+1} \sup_{[-\frac{\pi}{2}, \frac{\pi}{2}]} \|\nabla f(x)\|,$$

а общая ошибка не превосходит $\frac{\pi^n h \sqrt{n}}{2} \sup_{[-\frac{\pi}{2}, \frac{\pi}{2}]} \|\nabla f(x)\|$, то есть убывает линейно по h ,

однако высокая размерность требует для достижения высокой точности такого малого значения h , что вычисление квадратурной формулы с таким шагом не представляется возможным.

6.3 Результаты работы программы

Вычисление интеграла методом Монте–Карло. Из полученных оценок точности для того, чтобы получить результат с точностью $\varepsilon = 0.05$ и надежностью $1 - \alpha = 0.9$ необходимо провести

$$n = \frac{[f(Y)]}{\alpha \varepsilon^2} \approx 10^{10}$$

испытаний в методе Монте–Карло. Дисперсия здесь была оценена выборочной дисперсией, полученной на выборке размера $n = 10^7$. Интеграл вычисляется параллельно на 10^4 выборках размера 10^6 с последующим усреднением. Результат вычисления $J \approx 124.8225$.

Приведем также статистику десяти вычислений интеграла методом Монте–Карло на выборках размеров 10^6 , 10^7 , 10^8 (рис. 18). Доверительный интервал с уровнем доверия $\alpha = 0.95$ получен из неравенства Чебышева. Среднее время вычисления оценивается медианой.

Вычисление интеграла методом квадратур. С учетом всех оптимизаций, описанных в теоретической части, интеграл был вычислен на сетке размера 30 по каждому измерению, получен результат $J \approx 124.7040$, близкий к результату, полученному методом Монте–Карло.

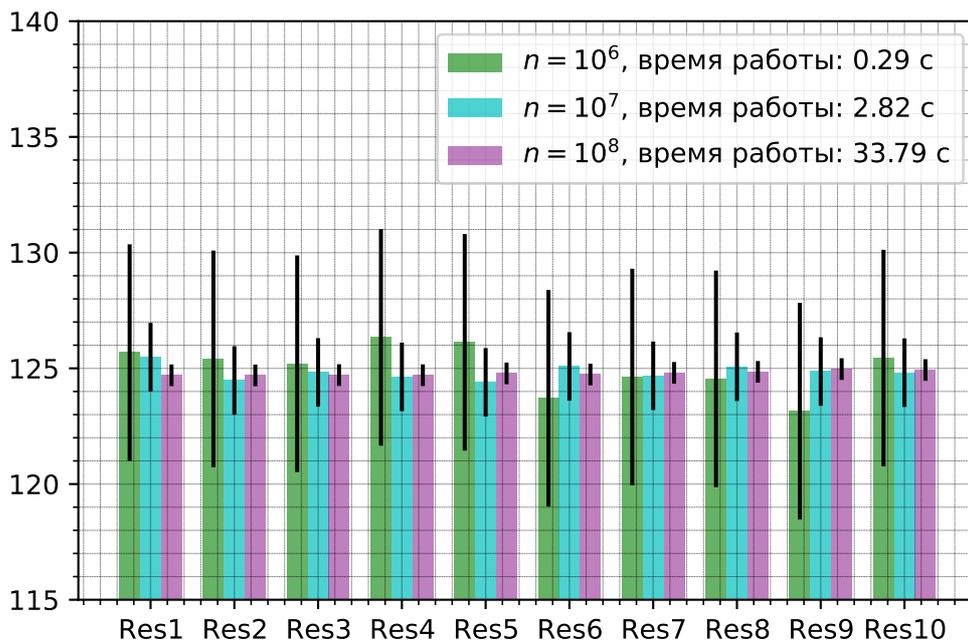


Рис. В.18: Точность метода Монте-Карло.

7 Задание 7

7.1 Постановка задачи

1. Методом случайного поиска найти минимальное значение функции f на множестве

$$A = \left\{ (x_1, x_2) : x_1^2 + x_2^2 \leq 1 \right\},$$

где

$$f(x_1, x_2) = x_1^3 \sin\left(\frac{1}{x_1}\right) + 10x_1x_2^4 \cos\left(\frac{1}{x_2}\right),$$

$f(0, 0) = 0$ по непрерывности.

2. Методом имитации отжига найти минимальное значение функции Розенброка g в пространстве \mathbb{R}^2 , где

$$g(x) = (x_1 - 1)^2 + 100(x_2 - x_1^2)^2.$$

3. Оценить точность. Сравнить результаты со стандартными методами оптимизации.

7.2 Теоретическая часть

Метод случайного поиска. Будем генерировать n равномерно распределенных по окружности A независимых случайных величин X_1, \dots, X_n и выбирать в качестве решения ту, на которой минимизируется значение f .

Найдем распределение полярных координат при равномерном распределении по окружности:

$$\mathbb{P}(X \in B) = \frac{1}{\pi} \iint_B dx dy = \frac{1}{\pi} \iint_B r dr d\varphi = \iint_B dr^2 d\left(\frac{\varphi}{2\pi}\right),$$

откуда можно сделать вывод, что $\varphi \sim \mathcal{U}(0, 2\pi)$, а r имеет функцию распределения

$$F_r(x) = \begin{cases} 0, & x < 0, \\ x^2, & x \in [0, 1], \\ 1, & x > 1. \end{cases}$$

Для моделирования r будем использовать метод обращения функции распределения:

$$r \stackrel{d}{=} \sqrt{U}, \quad U \sim \mathcal{U}[0, 1].$$

Точность метода случайного поиска. Оценим точность реализованного метода. Пусть $x^* = \operatorname{argmax}_{x \in A} f(x)$. Потребуем, чтобы для найденного решения x выполнялось $|f(x^*) - f(x)| < \varepsilon$ с вероятностью α . Так как функция f является непрерывно дифференцируемой в A , верна оценка

$$|f(x^*) - f(x)| \leq \|\nabla f(x^*)\| \|x - x^*\| \leq \sup_A \|\nabla f(x)\| \|x - x^*\| \leq \varepsilon$$

при $\|x - x^*\| \leq \delta = \frac{\varepsilon}{\sup_A \|\nabla f(x)\|}$.

Оценим величину $\sup_A \|\nabla f(x)\|$.

$$\frac{\partial f}{\partial x_1} = 3x_1^2 \sin\left(\frac{1}{x_1}\right) - x_1 \cos\left(\frac{1}{x_1}\right) + 10x_2^4 \cos\left(\frac{1}{x_2}\right),$$

$$\frac{\partial f}{\partial x_2} = 40x_1x_2^3 \cos\left(\frac{1}{x_2}\right) + 10x_1x_2^2 \sin\left(\frac{1}{x_2}\right),$$

$$\|\nabla f(x)\| = \sqrt{\left(\frac{\partial f}{\partial x_1}\right)^2 + \left(\frac{\partial f}{\partial x_2}\right)^2} \leq \sqrt{14^2 + 50^2} \leq 52.$$

Выберем число испытаний так, чтобы хотя бы один элемент выборки оказался в δ -окрестности x^* с вероятностью не менее α . Учтем также, что x^* может лежать на границе круга, и при малых δ окрестность $B_\delta(x^*) \cap A$ можно аппроксимировать полуокружностью.

$$\mathbb{P}(X_k \notin B_\delta(x^*), \quad k = 1, \dots, n) \leq \left(1 - \frac{\delta^2}{2}\right)^n \leq 1 - \alpha,$$

откуда $n = \frac{\ln(1 - \alpha)}{\ln\left(1 - \frac{\delta^2}{2}\right)}$.

Метод имитации отжига. Описание алгоритма.

0. Задаются входные параметры m, σ, t_0 и начальное значение (x_0, y_0) .
1. Генерируется кандидат на следующий шаг $(x_{k+1}, y_{k+1}) \sim (\mathcal{N}(x_k, t_k \sigma^2), \mathcal{N}(y_k, t_k \sigma^2))$ и вычисляется приращение функции $\Delta g = g(x_{k+1}, y_{k+1}) - g(x_k, y_k)$.
2. В случае, если $\Delta g < 0$, делается шаг в (x_{k+1}, y_{k+1}) . В противном случае переход выполняется с вероятностью $p_k = \frac{1}{2} \exp\left(-\frac{\Delta g}{t_k}\right)$.
3. Понижается температура по закону

$$t_{k+1} = t_k m^{\frac{1}{1+0.1|\Delta g|}}.$$

Такое правило позволяет медленнее понижать температуру в случае, когда начальное приближение далеко от решения, и приращения функции велики.

7.3 Результаты работы программы

Метод случайного поиска. Положим $\varepsilon = 0.01$, $\alpha = 0.99$. Тогда $\delta = \frac{0.01}{52}$, $n = 2.5 \cdot 10^8$. При таких параметрах результат работы метода

$$f_* \approx -1.2883, \quad x^* = (-0.3598, -0.9330).$$

Метод имитации отжига. Функция Розенброка достигает минимума в точке $x^* = (0, 0)$, где она принимает значение $f^* = 0$. Были подобраны следующие значения параметров в методе имитации отжига:

$$\sigma = 2.7, \quad m = 0.995, \quad t_0 = 2.$$

При таких значениях параметров алгоритм не совершает «фатальных» ошибок. Результат работы алгоритма, запущенного 1000 раз из случайной точки, координаты которой имеют распределение $\mathcal{N}(0, 100)$, показан на рис. 19. Все решения концентрируются на прямой, проходящей через точное решение. Приведем также некоторую статистику, относящуюся к работе метода:

- Среднее время работы: 0.116 с.
- Средняя частота переходов: 0.7%.
- Среднее число итераций: 10163.5.

Поскольку аналитическая оценка точности метода имитации отжига затруднительна, будем оценивать точность исходя из полученной выборки. На рис. 20 показаны распределения ошибки (расстояния от полученного результата до точного решения), а также ошибки по функционалу. Выборочные квантили порядка 0.99 равны 0.0209 и 0.0001 соответственно.

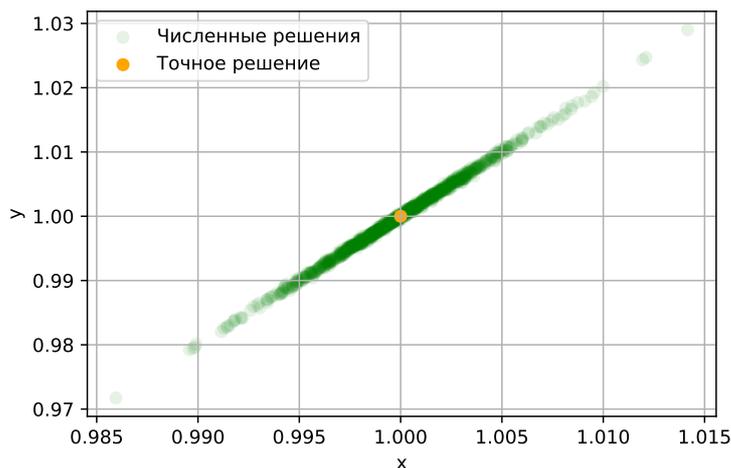
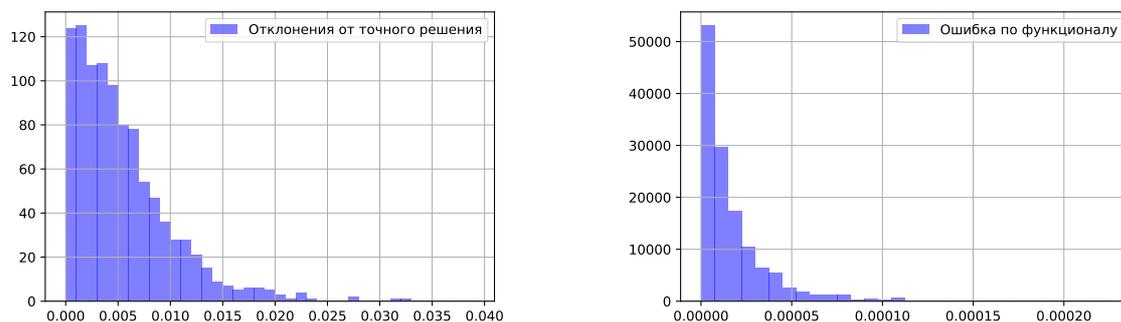


Рис. В.19: Результаты работы метода имитации отжига

Рис. В.20: Распределение величин $\|x - x^*\|$ и $\|f(x) - f(x^*)\|$.

Сравнение стохастических методов со стандартными методами оптимизации. Применим метод COBYLA (Constrained Optimization BY Linear Approximation) к задаче из пункта 1. При начальном приближении $x_0 = (-0.2, -0.8)$ получаем результат, похожий на результат работы метода случайного поиска:

$$f^* = -1.2885, \quad x^* = (-0.3574, -0.9330).$$

Однако если взять начальное приближение, далекое от точного решения, например, $x_0 = (0.2, 0.1)$, метод дает неверный результат

$$f^* = -0.0122, \quad x^* = (0.2455, -0.1110).$$

Метод сопряженных градиентов, выпущенный из $x_0 = (5, 5)$, дает точное решение задачи минимизации функции Розенброка:

$$f^* = 2 \cdot 10^{-12}, \quad x^* = (0.9999968, 0.9999972).$$

8 Задание 8

8.1 Постановка задачи

1. Применить метод Монте–Карло к решению первой краевой задачи для двумерного уравнения Лапласа в единичном круге:

$$\begin{cases} \Delta u = 0, & (x, y) \in D = \{(x, y): x^2 + y^2 \leq 1\} \\ u|_{\partial D} = f(x, y) \\ u \in C^2(D), f \in C(\partial D) \end{cases}$$

Для функции $f(x, y) = x^2 - y^2$ найти аналитическое решение и сравнить с полученным по методу Монте–Карло.

8.2 Теоретическая часть

Применение метода Монте–Карло. Введем на плоскости сетку $x_{ij} = (-1+ih, -1+jh)$ с шагом h и назовем внутренними узлы, для которых все 4 соседних узла лежат в круге. Узлы, лежащие в круге, но не удовлетворяющие этому условию, будем называть крайними.

Аппроксимируя вторые частные производные, можно свести задачу Дирихле к следующей разностной схеме:

$$u_{ij} = \frac{1}{4}(u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1}) \text{ для внутренних узлов,}$$

$$u_{ij} = f_{ij} \text{ для крайних узлов.}$$

Рассмотрим теперь следующую модель случайного блуждания:

1. Частица начинает свой путь во внутреннем узле x_{ij} .
2. С равными вероятностями она переходит в один из четырех соседних узлов.
3. Процесс продолжается до тех пор, пока частица не попадет на крайний узел x_{pq} .

Обозначим $p(x_{ij}, x_{pq})$ вероятность, начав путь в узле x_{ij} , закончить его в узле x_{pq} . Из формулы полной вероятности следует

$$p(x_{ij}, x_{pq}) = \frac{1}{4}(p(x_{i-1,j}, x_{pq}) + p(x_{i+1,j}, x_{pq}) + p(x_{i,j-1}, x_{pq}) + p(x_{i,j+1}, x_{pq})) \quad (\text{B.1})$$

для внутренних узлов x_{ij} и $p(x_{ij}, x_{ij}) = 1$ для крайних узлов.

Рассмотрим теперь случайную величину $\xi_{ij} = f(x_{pq})$ — значение f в конечной точке пути. Так как $\mathbb{E}\xi_{ij} = \sum_{p,q} p(x_{ij}, x_{pq})f(x_{pq})$, используя формулу (B.1) связи вероятностей на соседних узлах, получим

$$\mathbb{E}\xi_{ij} = \frac{1}{4}(\mathbb{E}\xi_{i-1,j} + \mathbb{E}\xi_{i+1,j} + \mathbb{E}\xi_{i,j-1} + \mathbb{E}\xi_{i,j+1}) \text{ для внутренних узлов,}$$

$$\mathbb{E}\xi_{ij} = f_{ij} \text{ для крайних.}$$

Видно, что эти уравнения совпадают с разностной схемой, аппроксимирующей исходную задачу.

Значения $\mathbb{E}\xi_{ij}$ в каждом узле можно вычислить методом Монте–Карло, моделируя описанный выше случайных процесс. Данный метод подробно описан в [6].

Для оптимизации вычислений воспользуемся тем, что моделируемый процесс случайного блуждания обладает марковским свойством, поэтому при прохождении траектории через промежуточные узлы дальнейший путь частицы можно рассматривать как процесс, начавшийся в этом промежуточном узле.

8.3 Результаты работы программы

На рис. 21 показано решение, полученное методом Монте–Карло на сетке с шагом $h = 0.05$ при проведении $n = 10^4$ испытаний в каждой точке, а также аналитическое решение $f(x, y) = x^2 - y^2$. Максимальное отклонение от аналитического решения равно 0.0484. Разность численного и аналитического решения показана на рис. 22.

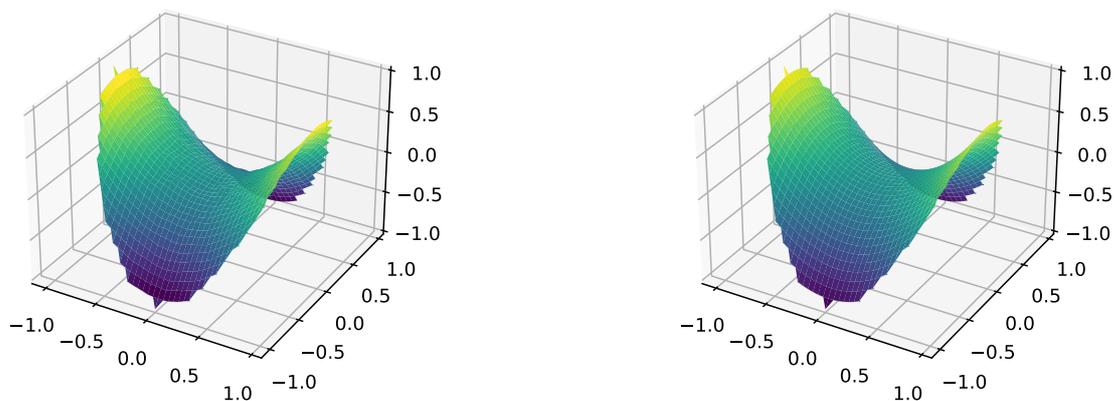


Рис. В.21: Решение, полученное методом Монте–Карло, и аналитическое решение.

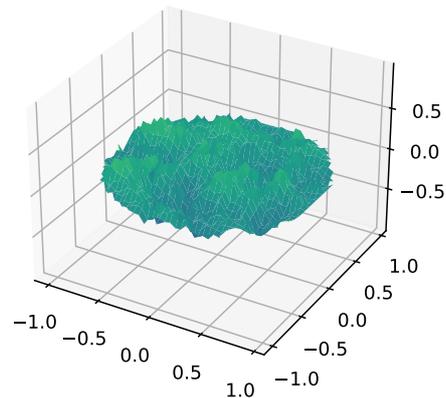


Рис. В.22: Разность численного и аналитического решения при $n = 10^4$.

9 Задание 9

9.1 Постановка задачи

Рассмотреть два вида процессов:

- Винеровский процесс $W(t), t \in [0, 1], W(0) = 0$.
- Процесс Орнштейна–Уленбека $X(t), t \in [0, 1], X(0) = X_0$, то есть стационарный марковский гауссовский процесс. Начальное значение X_0 генерируется случайным образом так, чтобы полученный процесс бы стационарным.

Для данных процессов

1. Найти ковариационную функцию и переходные вероятности.
2. Промоделировать независимые траектории процесса с данными переходными вероятностями методом добавления разбиения отрезка.
3. Построить график траекторий.

9.2 Теоретическая часть

Метод разбиения отрезка. Опишем кратко метод добавления разбиения отрезка. Пусть требуется смоделировать случайный процесс $X(t)$ на отрезке $[0, 1]$.

1. Моделирование $X_0 = X(0)$.
2. Моделирование $X(1)$ по условному распределению $[X(1)|X(0)]$.

3. Между каждыми двумя соседними узлами t_i, t_{i+1} сетки добавим узел $t = \frac{t_i + t_{i+1}}{2}$ и смоделируем процесс в нем, используя условное распределение случайной величины

$$[X(t)|X(t_i) = x_i, X(t_{i+1}) = x_{i+1}].$$

4. Повторение пункта 3 до тех пор, пока шаг сетки не станет достаточно мелким.

Винеровский процесс. Винеровский процесс — гауссовский процесс с независимыми приращениями такой, что

1. $W(0) \stackrel{\text{п.п.}}{=} 0$
2. $W(t+h) - W(t) \sim \mathcal{N}(0, \sigma^2 h)$

Найдем его ковариационную функцию функцию, предполагая $t_1 < t_2$:

$$\begin{aligned} R(t_1, t_2) &= \mathbb{E}W(t_1)W(t_2) = \mathbb{E}W^2(t_1) + \mathbb{E}[W(t_1) - W(0)][W(t_2) - W(t_1)] = \\ &= \sigma^2 t_1 + \mathbb{E}W(t_1)\mathbb{E}[W(t_2) - W(t_1)] = \sigma^2 t_1, \end{aligned}$$

откуда

$$K(t_1, t_2) = \min(t_1, t_2)\sigma^2.$$

По определению $W(0) = 0$, $[W(1)|W(0)] \sim \mathcal{N}(0, \sigma^2)$.

Найдем теперь для $t_1 < t_2 < t_3$ распределение $Y = [W(t_2)|W(t_1) = x_1, W(t_3) = x_3]$. Так как $W(t)$ имеет нормальное распределение с параметрами 0 и $\sigma^2 t$, Y имеет плотность

$$p_Y(x_2) = \frac{p_{123}(x_1, x_2, x_3)}{p_{13}(x_1, x_3)},$$

где p_{123} — совместная плотность $(W(t_1), W(t_2), W(t_3))$, p_{13} — совместная плотность вектора $(W(t_1), W(t_3))$.

Известно, что плотность n -мерного нормального распределения с матожиданием μ и ковариационной матрицей Σ имеет вид

$$p(x) = \frac{1}{\sqrt{(2\pi)^n |\Sigma|}} \exp\left(-\frac{1}{2}(x - \mu)^T \Sigma^{-1}(x - \mu)\right) \quad (\text{B.2})$$

Используя полученную для ковариационной функции формулу, получаем

$$\begin{aligned} \Sigma_{13} &= \sigma^2 \begin{pmatrix} t_1 & t_1 \\ t_1 & t_3 \end{pmatrix}, \\ \Sigma_{123} &= \sigma^2 \begin{pmatrix} t_1 & t_1 & t_1 \\ t_1 & t_2 & t_2 \\ t_2 & t_2 & t_3 \end{pmatrix}. \end{aligned}$$

С помощью библиотеки для символьных вычислений SymPy найдем

$$p_Y(x_2) = C \exp \left\{ -\frac{1}{2} \frac{4}{t_3 - t_1} \left(x_2 - \frac{x_1 + x_3}{2} \right)^2 \right\},$$

то есть, обозначая шаг измельченной сетки за $h = t_2 - t_1$, получим

$$Y \sim \mathcal{N} \left(\frac{x_1 + x_3}{2}, \frac{h\sigma^2}{2} \right).$$

Процесс Орнштейна–Уленбека. Процесс Орнштейна–Уленбека — это стационарный марковский гауссовский процесс. Известно [8], что для такого процесса

$$\rho(X(t_1), X(t_3)) = \rho(X(t_1), X(t_2))\rho(X(t_2), X(t_3)), \quad t_1 < t_2 < t_3,$$

где ρ — коэффициент корреляции.

В силу однородности процесса для любого t верно $X(t) \sim \mathcal{N}(0, \sigma^2)$ (без ограничения общности будем рассматривать центрированный процесс). Найдем вид функции $K(t) = \rho(X(t), X_0)$, предполагая $0 \leq t_1 < t_2 \leq 1$:

$$K(t_2) = \rho(X(t_2), X_0) = \rho(X(t_2), X(t_1))\rho(X(t_1), X_0) = K(t_2 - t_1)K(t_1).$$

То есть для любых $t, s \geq 0$ верно $K(t + s) = K(t)K(s)$. Логарифм этой функции $f(x) = \ln K(x)$ является решением уравнения Коши

$$f(x + y) = f(x) + f(y),$$

и, так как функция $K(t)$ непрерывна, она имеет вид $K(t) = e^{-\lambda t}$, $\lambda > 0$, а ковариационная функция процесса

$$R(t_1, t_2) = \sigma^2 e^{-\lambda|t_2 - t_1|}, \quad \lambda > 0.$$

Сгенерируем X_0 как нормальную случайную величину с нулевым матожиданием и дисперсией σ^2 . Найдем теперь распределение $[X(1)|X_0]$. Ковариационная матрица для $(X_0, X(1))$ имеет вид

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & e^{-\lambda} \\ e^{-\lambda} & 1 \end{pmatrix}.$$

Проведя необходимые расчеты, получаем

$$[X(1)|X_0] \sim \mathcal{N} \left(x_0 e^{-\lambda}, \sigma^2 (1 - e^{-2\lambda}) \right).$$

Найдем теперь распределение $Y = [X(t_2)|X(t_1) = x_1, X(t_3) = x_3]$, $t_1 < t_2 < t_3$.

Обозначим шаг новой сетки за $h = t_2 - t_1 = t_3 - t_2$ и, как и в случае винеровского процесса, найдем ковариационные матрицы

$$\Sigma_{13} = \sigma^2 \begin{pmatrix} 1 & e^{-2\lambda h} \\ e^{-2\lambda h} & 1 \end{pmatrix},$$

$$\Sigma_{123} = \sigma^2 \begin{pmatrix} 1 & e^{-\lambda h} & e^{-2\lambda h} \\ e^{-\lambda h} & 1 & e^{-\lambda h} \\ e^{-2\lambda h} & e^{-\lambda h} & 1 \end{pmatrix}.$$

После преобразований находим, что

$$Y \sim \mathcal{N} \left(\frac{x_1 + x_3}{e^{\lambda h} + e^{-\lambda h}}, \frac{\sigma^2(e^{2\lambda h} - 1)}{e^{2\lambda h} + 1} \right).$$

9.3 Результаты работы программы

Все процессы смоделированы на сетке размера $N = 2^{15} + 1 \approx 3.3 \cdot 10^4$. Из ковариационной функции процесса Орнштейна–Уленбека видно, что с увеличением λ корреляция между значениями быстро уменьшается, и процесс становится похож на белый шум. При $\lambda = 0$ корреляция между любыми двумя значениями равна 1, и процесс является постоянной функцией $X_t \equiv X_0$.



Рис. В.23: Моделирование стандартного винеровского процесса W_t .



Рис. В.24: Моделирование процесса Орнштейна–Уленбека с $\lambda = 0.1$, $\sigma = 1$.



Рис. В.25: Моделирование процесса Орнштейна–Уленбека с $\lambda = 10$, $\sigma = 1$.

10 Задание 10

10.1 Постановка задачи

Произвести фильтрацию одномерного процесса Орнштейна–Уленбека:

1. Используя генератор белого шума, добавить случайную ошибку с известной дисперсией к реализации процесса Орнштейна–Уленбека.
2. При помощи одномерного фильтра Калмана оценить траекторию процесса по зашумленному сигналу. Параметры процесса и белого шума считать известными.
3. Рассмотреть случай, когда шум
 - является гауссовским;
 - имеет распределение Коши.

10.2 Теоретическая часть

Алгоритм работы фильтра Калмана. Рассмотрим сетку

$$\{t_n = nh, n = 0, 1, \dots, N\}, \quad h = \frac{1}{N}$$

и значения процесса Орнштейна-Уленбека $x_n = X(t_n)$.

Представим процесс в виде дискретной динамической системы вида

$$x_{n+1} = ax_n + \nu_n, \quad \text{i.i.d. } \nu_n \sim \mathcal{N}(0, q), \quad x_1 \sim \mathcal{N}(0, \sigma^2).$$

Обозначим зашумленный сигнал

$$y_n = x_n + \varepsilon_n, \quad \text{i.i.d. } \varepsilon_n \sim \mathcal{N}(0, r).$$

Параметры σ, λ процесса считаем известными. Найдем через них параметры системы a, q . Для этого запишем ковариационную функцию для процесса в моменты t_n и t_{n+1} :

$$\begin{aligned} R(t_n, t_n) &= (x_n) = \sigma^2, \\ R(t_n, t_{n+1}) &= \text{cov}(x_n, x_{n+1}) = a(x_n) = a\sigma^2 = \sigma^2 e^{-\lambda h}, \\ R(t_{n+1}, t_{n+1}) &= \text{var}(x_{n+1}) = a^2(x_n) + q = a^2\sigma^2 + q = \sigma^2. \end{aligned}$$

Отсюда легко выразить

$$a = e^{-\lambda h}, \quad q = \sigma^2(1 - e^{-2\lambda h}).$$

Опишем один шаг дискретного фильтра Калмана. Обозначим за $\hat{x}_{n|n}$ оценку значения x_n при известных y_1, \dots, y_n . $\hat{x}_{n|n-1}$ — экстраполяция процесса на следующий шаг в соответствии с динамической системой. Через $p_{n|n}$ будем обозначать дисперсию ошибки фильтрации на n -м шаге, через $p_{n|n-1}$ — прогнозируемую на следующем шаге дисперсию.

1. Прогнозируем значение процесса и дисперсию ошибки

$$\hat{x}_{n|n-1} = a\hat{x}_{n-1|n-1}, \quad p_{n|n-1} = a^2 p_{n-1|n-1} + q.$$

2. Вычисляем разницу между наблюдаемым процессом и прогнозом и коэффициент усиления Калмана

$$\delta_n = y_n - \hat{x}_{n|n-1}, \quad k_n = \frac{p_{n|n-1}}{p_{n|n-1} + r}.$$

3. В качестве результата фильтрации берем выпуклую комбинацию наблюдаемого и предсказанного значения

$$\hat{x}_{n|n} = \hat{x}_{n|n-1} + k_n \delta_n, \quad p_{k|k} = (1 - k_n) p_{n|n-1}.$$

Дискретный фильтр Калмана минимизирует среднеквадратическую ошибку в классе линейных оценок. Доказательство этого факта можно найти, например, в [7].

Поскольку рассматривается линейный фильтр гауссовского процесса с гауссовским шумом, ошибка фильтрации будет иметь нормальное распределение, а доверительный интервал с уровнем доверия α будет иметь вид

$$[\hat{x}_{n|n} - \Delta, \hat{x}_{n|n} + \Delta], \quad \Delta = -\sqrt{p_{n|n}} \Phi^{-1} \left(\frac{\alpha}{2} \right).$$

10.3 Результаты работы программы

Нормально распределенная ошибка. Всяду в этом задании рассматриваем процесс Орнштейна–Уленбека с $\lambda = 0.01$, $\sigma = 1$. Добавим к нему нормальный шум с дисперсией $r = 0.01$. (рис. 26). Результат фильтрации показан на рис. 27, а доверительные интервалы с уровнем доверия $\alpha = 0.95$ на рис. 28. При этом вне интервалов процесс оказывается в 3.19% времени.

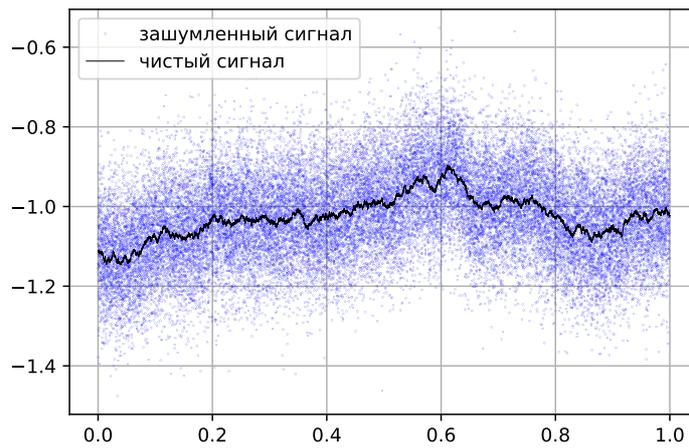
Рис. В.26: Ошибка распределена нормально, $r = 0.01$.

Рис. В.27: Результат работы фильтра Калмана.

Ошибка имеет распределение Коши. При добавлении шума, имеющего распределение Коши с нулевым параметром сдвига и с параметром масштаба, равным 0.0001, в зашумленном сигнале появляются сильные выбросы, возникающие из-за тяжелых хвостов распределения Коши (рис. 29). Эти же выбросы иногда сохраняются и после фильтрации (рис. 30), и в доверительных интервалах (рис. 31). Вне доверительного интервала порядка $\alpha = 0.95$ оказывается 1.71% траектории.



Рис. В.28: Доверительные интервалы.

Рис. В.29: Ошибка имеет распределение Коши с параметром масштаба 10^{-4} .



Рис. В.30: Результат работы фильтра Калмана.



Рис. В.31: Доверительные интервалы.

11 Задание 11

11.1 Постановка задачи

Построить двумерное пуассоновское поле, отвечающее сложному пуассоновскому процессу:

1. Первая интерпретация: система массового обслуживания. При этом, первая координата поля — время поступления заявки в СМО (равномерное распределение), вторая — время ее обслуживания (распределение χ^2 с 10 степенями свободы).
2. Вторая интерпретация: система массового обслуживания с циклической активностью $\lambda(1 + \cos(t))$ и единичными скачками. Свести данную задачу при помощи метода Льюиса и Шадлеара к моделированию двумерного пуассоновского поля, где первая координата имеет равномерное распределение, а вторая — распределение Бернулли.

3. Третья интерпретация: работа страховой компании. Первая координата — момент наступления страхового случая (равномерное распределение), вторая координата — величина ущерба (распределение Парето). Поступление капитала по времени линейно со скоростью $c > 0$, начальный капитал $W > 0$.
4. Для каждой системы рассмотреть все возможные случаи поведения системы в зависимости от значения параметров.

11.2 Теоретическая часть

Моделирование СМО. Будем моделировать систему массового обслуживания на отрезке времени $[0, T]$. Обозначим $\eta \sim \text{Pois}(\lambda T)$ число заявок, поступивших в систему за время от 0 до T . λ — интенсивность поступления заявок, то есть среднее число новых заявок в единицу времени.

Пусть t_1, \dots, t_η — времена поступления заявок, $t_j \sim \mathcal{U}[0, T]$, $t_{(j)}$ — j -я порядковая статистика, $s_j \sim \chi^2(10)$ — время обработки j -й заявки (то есть заявки, поступившей в момент $t_{(j)}$).

Обозначим X_j время окончания обработки j -й заявки. Заметим, что

$$X_1 = t_{(1)} + s_1, \quad X_j = \begin{cases} t_{(j)} + s_j, & X_{j-1} \leq t_{(j)} \\ X_{j-1} + s_j, & X_{j-1} > t_{(j)} \end{cases}$$

Таким образом, число заявок в очереди в момент времени t можно выразить следующим образом:

$$N(t) = \sum_{j=1}^{\eta} \mathbb{1}_{t_{(j)} \leq t < X_j},$$

то есть как число заявок, которые уже поступили к моменту t , но еще не были обработаны.

Будем моделировать $\eta \sim \text{Pois}(\lambda T)$, затем $t_1, \dots, t_\eta \sim \mathcal{U}[0, 1]$, $s_1, \dots, s_\eta \sim \chi^2(10)$ как независимые случайные величины. Затем на основе полученной выборки вычисляются X_1, \dots, X_η и $N(t)$.

Заметим также, что время между заявками имеет экспоненциальное распределение с параметром λ , то есть среднее время между заявками равно $\frac{1}{\lambda}$. Среднее время обработки одной заявки равно $\mathbb{E}s_1 = 10$. Таким образом, если $\frac{1}{\lambda} < 10$, то есть $\lambda > \frac{1}{10}$, то заявки поступают с большей скоростью, чем успевают обрабатываться, и длина очереди в среднем растет. Если $\lambda < \frac{1}{10}$, то скорость обработки превышает скорость поступления. $\lambda = \frac{1}{10}$ соответствует некоторому промежуточному состоянию, при котором скорость поступления заявок совпадает со скоростью их обработки.

Моделирование СМО с циклической активностью. Для моделирования СМО с циклической интенсивностью $\lambda(t) = \lambda(1 + \cos(t))$ будем использовать метод Льюиса и Шедлеара (его обоснование можно найти в [3]). Так как $\lambda(t) \leq 2\lambda$ для любых t , будем использовать однородный пуассоновский процесс с интенсивностью $\lambda^* = 2\lambda$. Опишем алгоритм моделирования.

1. Генерируем $\eta \sim \text{Pois}(\lambda^*)$ — число поступивших заявок — и $t_1^*, \dots, t_\eta^* \sim \mathcal{U}[0, T]$ — времена поступления заявок в процессе $N^*(\cdot)$.
2. Для каждого j удаляем t_j^* из выборки с вероятностью $1 - \frac{\lambda(t_j^*)}{\lambda^*(t_j^*)}$.
3. Оставшиеся точки являются моментами поступления заявки в процессе $N(\cdot)$.

Так как средняя мгновенная интенсивность процесса равна λ , сценарии функционирования СМО аналогичны пункту 1.

Моделирование работы страховой компании. Пусть число страховых случаев к моменту t равно $N(t)$ — однородный пуассоновский процесс с интенсивностью λ . Тогда капитал страховой компании изменяется по закону

$$K(t) = W + ct - \sum_{k=1}^{N(t)} \xi_k,$$

где ξ_k имеет распределение Парето с плотностью $p_\xi(x) = \frac{k}{x^{k+1}}$, $x > 1$. Параметр распределения Парето $k > 0$.

Если в какой-то момент τ капитал $K(\tau)$ достигает нуля, то $K(t) = 0$, $t > \tau$.

Алгоритм моделирования аналогичен предыдущим пунктам: генерируется $\eta \sim \text{Pois}(\lambda T)$, $t_1, \dots, t_\eta \sim \mathcal{U}[0, T]$. Случайные величины ξ_1, \dots, ξ_η , имеющие распределение Парето с параметром k , будем моделировать методом обращения функции распределения:

$$\xi_j = (1 - U_j)^{-\frac{1}{k}}, \quad U_j \sim \mathcal{U}[0, 1], \quad j = 1, \dots, \eta.$$

Далее по вышеописанным формулам вычисляется капитал $K(t)$.

При $k > 1$ распределение Парето имеет математическое ожидание, равное $\frac{k}{k-1}$. Тогда легко показать, что

$$\mathbb{E}K(t) = W + ct - \lambda t \frac{k}{k-1},$$

то есть при $c - \frac{\lambda k}{k-1} > 0$ капитал компании в среднем будет расти. При противоположном неравенстве $K(t)$ в среднем будет убывать, приводя к разорению.

При $k \leq 1$ математическое ожидание ξ_j равно бесконечности, и на достаточно большом промежутке времени компания будет разоряться при любых значениях $c, W, \lambda > 0$.

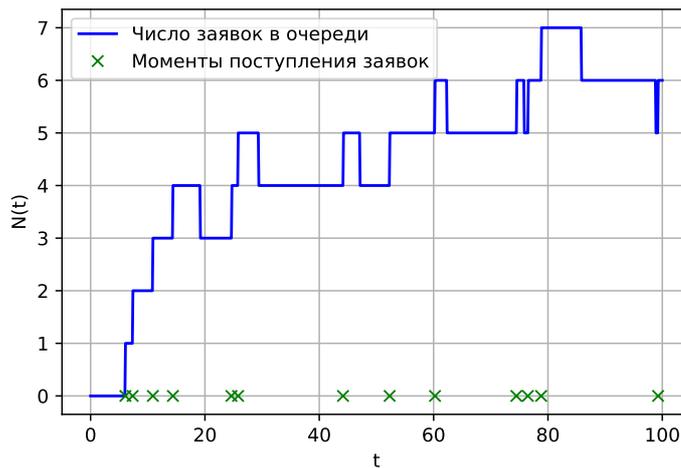


Рис. В.32: Работа СМО с $\lambda = 0.1$

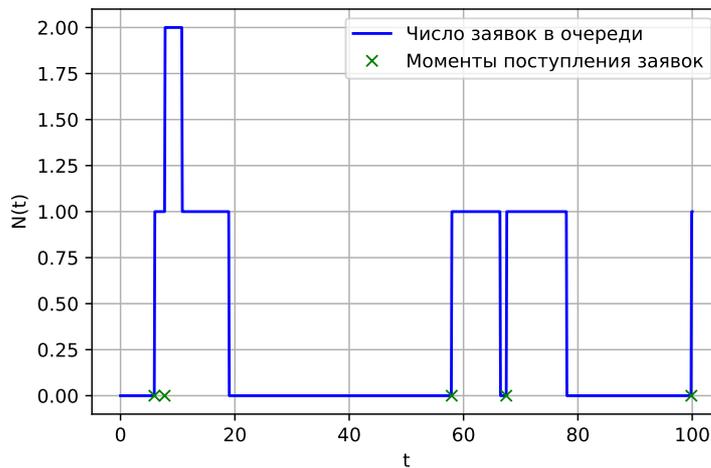


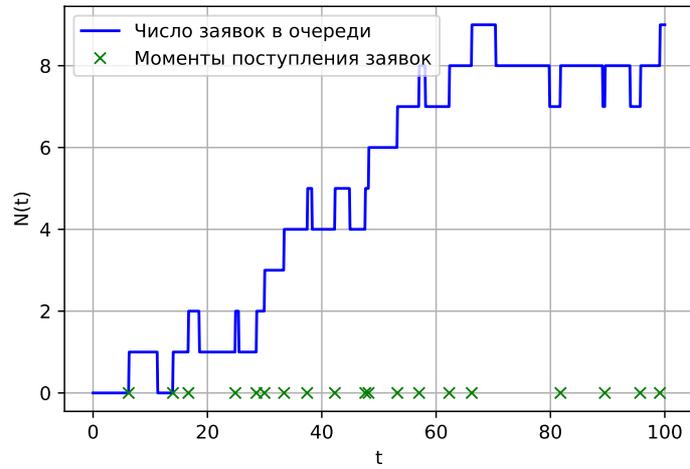
Рис. В.33: Работа СМО с $\lambda = 0.05$

11.3 Результаты работы программы

Моделирование СМО. На рис. 32-34 показаны результаты моделирования СМО с параметрами $\lambda = 0.1, 0.05, 0.2$ соответственно, что соответствует трем основным сценариям, описанным в теоретической части.

На рис. 35 продемонстрирована работа СМО с циклической интенсивностью на отрезке $[0, 4\pi]$ с параметром $\lambda = 10$. Видно, что в моменты, когда интенсивность близка к нулю, новых заявок практически не поступает.

Моделирование работы страховой компании. На рис. 36-38 показаны сценарии работы страховой компании на отрезке $[0, 100]$. Параметры модели $W = 30, k = 2, c = 2$. Параметр λ изменяется так, чтобы охватить все три описанных в теоретической части случая.

Рис. В.34: Работа СМО с $\lambda = 0.2$ Рис. В.35: Работа СМО с циклической интенсивностью. $\lambda = 10$

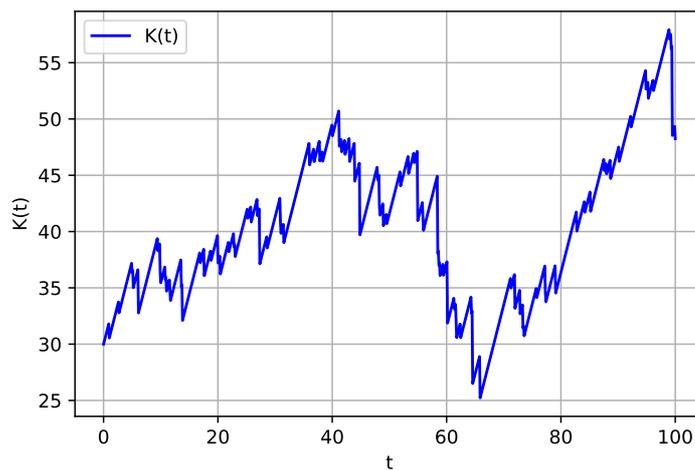


Рис. В.36: Моделирование работы страховой компании с $W = 30$, $k = 2$, $c = 2$, $\lambda = 1$.

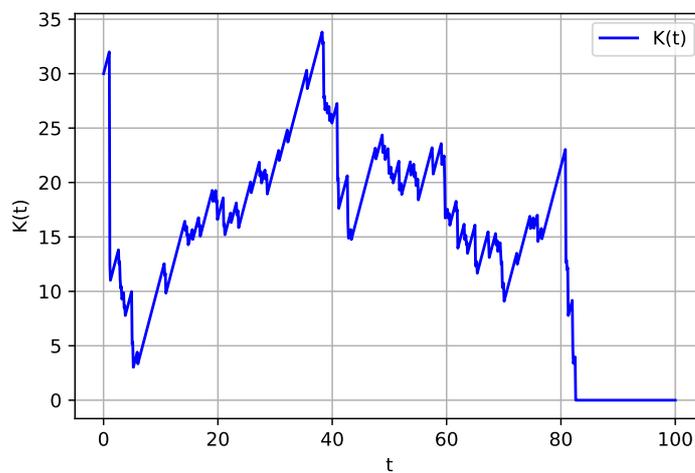


Рис. В.37: Моделирование работы страховой компании с $W = 30$, $k = 2$, $c = 2$, $\lambda = 1.1$.

На рис. 39 показано, что при $k = 1$, когда распределение Парето не имеет математического ожидания, один страховой случай может полностью разорить фирму.

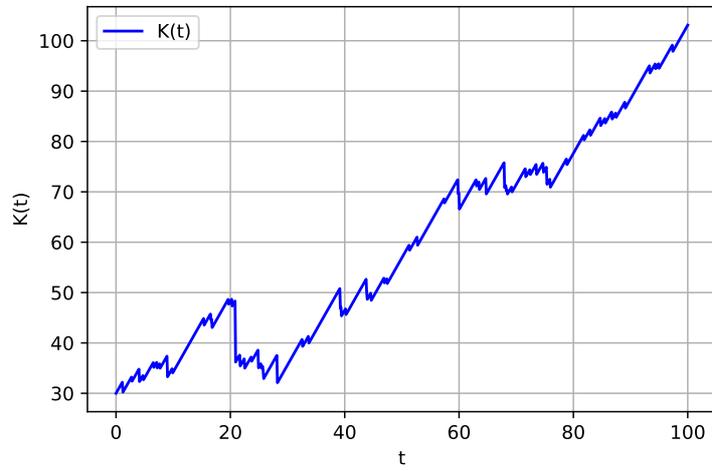


Рис. В.38: Моделирование работы страховой компании с $W = 30$, $k = 2$, $c = 2$, $\lambda = 0.9$.

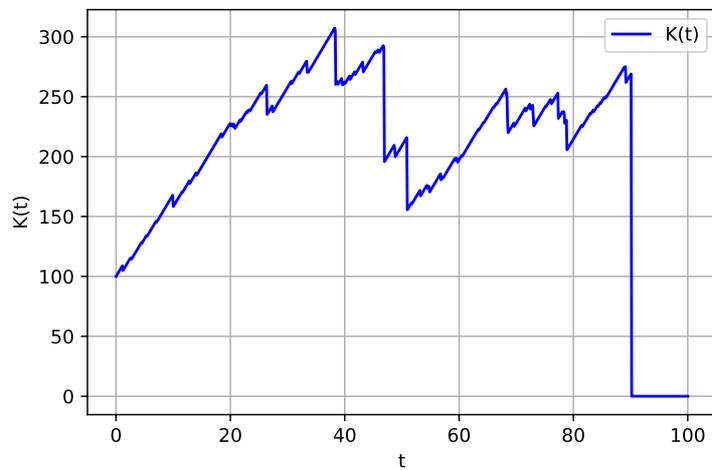


Рис. В.39: Моделирование работы страховой компании с $W = 100$, $k = 1$, $c = 8$, $\lambda = 1$.

Литература

- [1] Смирнов, С.Н., Лекции по курсу «Стохастический анализ и моделирование», 2020.
- [2] DeGroot, М.Н. & М.Ј. Schervish (2011), Probability and Statistics, 4th Ed, Pearson.
- [3] Lewis, P. A. W., & Shedler, G. S. (1979). Simulation of nonhomogeneous poisson processes by thinning. *Naval Research Logistics Quarterly*, 26(3), 403–413. doi:10.1002/nav.3800260304
- [4] Kolmogorov A (1933), Sulla determinazione empirica di una legge di distribuzione. *G. Ist. Ital. Attuari.* 4: 83–91.
- [5] Smirnov, N. (1939), On the estimation of the discrepancy between empirical curves of distribution for two independent samples. *Bull. Math. Univ. Moscou* 2:2.
- [6] Бусленко Н.П., Голенко Д.И., Соболев И.М., Срагович В.Г., Шрейдер Ю.А. Метод статистических испытаний. 1962. 332 с. Букинист.
- [7] Острем К.Ю. Введение в стохастическую теорию управления. М.: Мир, 1973. — 324 с.
- [8] Феллер В. Введение в теорию вероятностей и ее приложения (том 2). М.: Мир, 1967 г. — 765 с.
- [9] Ширяев А. Н. Вероятность, — М.: Наука. 1989.